Early Response-to-Intervention Measures and Criteria as Predictors of Reading Disability in the Beginning of Third Grade

Kristen D. Beach

Rollanda E O'Connor

Beach, K.D., & O'Connor, R.E. (2015). Early Response-to-Intervention Measures and Criteria as Predictors of Reading Disability in the Beginning of Third Grade. *Journal of Learning Disabilities, 48,* 196-223.

The research in this article was supported by Grant R324B070098 from the U.S. Department of Education, Institute of Education Sciences to the University of California, Riverside. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Department of Education, Institute of Education Sciences.

Abstract:

We explored the usefulness of 1st and 2nd grade reading measures and responsiveness criteria collected within a Response-to-Intervention (RtI) framework for predicting Reading Disability (RD) in 3rd grade. We used existing data from 387 linguistically diverse students who had participated in a longitudinal RtI study. Model-based predictors of RD were analyzed using logistic regression; isolated measure/criteria combinations for predicting RD were analyzed using classification analysis. Models yielded superior classification rates compared to single measure approaches, and did not systematically misclassify English Learners. However, particular 1st and 2nd grade measure/criteria combinations also showed promise as isolated predictors of RD in word reading/text fluency. Model based approaches were required for acceptable classification of students with RD in comprehension. While the former finding is promising for early identification of students in need of more intensive instruction in lexical or fluency-based skills, the latter finding reaffirms literature attesting to the complexity of RD in comprehension and difficulty of predicting deficits using early measures of reading, which primarily assess word reading skill. Results replicated well with an independent sample, thus enhancing confidence in study conclusions. Implications regarding the use of RtI for predicting RD are discussed.

Early Response-to-Intervention Measures and Criteria as Predictors of Reading Disability in the Beginning of Third Grade.

Over a decade ago, researchers voiced concerns regarding the use of the IQ-Discrepancy model for identifying learning disabilities (e.g. see Vellutino et al., 1996; Vellutino, Scanlon, & Lyon, 2000) and empirical support for alternative disability identification methods arose shortly thereafter (see Speece & Case, 2001). The reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA; P.L. 108-446) in 2004 allowed schools' use of Response to Intervention (RtI) as a framework for special education referral; thus exploring the conditions under which students' poor response to intervention is a valid indicator of learning disabilities became a prominent focus in special education research. Recent reports suggest that alternative disability identification methods, such as those used within RtI, can accurately predict the development of reading disabilities for struggling readers receiving Tier II intervention. However, studies often use different measures and/or criteria to assess students' progress during intervention and proficiency after intervention, which precludes consensus on measures and criteria that optimally classify students. Additionally, few researchers have attempted to predict reading achievement for struggling readers past 2nd grade and none have included struggling readers who began school reading on target, but fell behind after three or more years of instruction. Addressing these limitations is essential if RtI methods are adopted to identify children who are or will be in need of special education services.

Although research supports RtI for early intervention and prevention efforts (O'Connor, 2000; Torgesen, 2000; Vaughn, Linan-Thompson, & Hickman, 2003; Wanzek & Vaughn, 2007), researchers are reluctant to recommend RtI as a disability identification tool. One hesitation

stems from variability in measures and criteria used by researchers (Barth, Stuebing, Anthony, Denton, Mathes, Fletcher, & Francis, 2008; D. Fuchs, L.S. Fuchs, & Compton, 2004; Fuchs, Compton, Fuchs, Bryant & Davis, 2008) and schools (Mellard, McKnight, & Woods, 2009) to identify good and poor responders during Tier I and II instruction, as well as average and disabled readers after intervention or at some later time. Currently, researchers and schools (see Mellard et al., 2009) report using Curriculum Based Measures (CBM) such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2003), Word Identification Fluency measures (WIF; Fuchs et al., 2004; Zumeta, Compton, & Fuchs, 2012), and standardized measures including the Woodcock Reading Mastery Tests (WRMT; Woodcock, 1998) and Gray Oral Reading Tests (GORT, Wiederholt & Bryant, 2001) to judge responsiveness to intervention and final reading proficiency or reading disability (RD). Because distinct groups of students are identified as good responders, poor responders, average readers, and RD readers depending on the measure selected to gauge responsiveness and proficiency (e.g. Barth et al., 2008; Fuchs et al., 2004; 2008; Simmons, Coyne, Kwok, McDonagh, Harn & Kame'enui 2008), it is important to determine whether particular measurement tools provide more reliable RD classifications, either when used in combination or isolation. This was the first goal of the present study.

In addition to variability among measures, criteria selection (i.e. the method of determining the score(s) below which students are classified as poor responders) also varies across studies and school sites. Criteria often used to determine intervention responsiveness include: low growth, dual discrepancy, median split, final benchmark, low final achievement, and final normalization (see O'Connor & Klingner, 2010; Fuchs & Fuchs, 2006; and Waesche, Schatschneider, Maner, Ahmed, & Wagner, 2001 for descriptions). Like measure variability,

criteria variability leads to disagreements on students' membership in good/poor responder and average/RD reader groups (Barth et al., 2008; Burns & Senesac, 2005; Fuchs et. al 2004; 2008). Some researchers (e.g. Francis, Fletcher, Stuebing, Lyon, Shaywitz, & Shaywitz, 2005) argue that applying any cut-point to a continuous measure results in unstable group membership, especially when behavioral factors (including responsiveness to intervention) are not considered. Nevertheless, others (Fuchs et al., 2008; Speece & Case, 2001) have found particular criteria useful for identifying students who need more intensive instruction.

Stability of RD designations made in 1st and 2nd grade is an additional concern; if RtI methods are used to identify students in need of intensive instruction and/or special education in early grades, those decisions must be made with confidence that continued instruction or Tier II intervention will not change a student's trajectory. Students might be identified for more intensive (e.g. Tier III) intervention or begin special education eligibility processes in 1st or 2nd grade if RtI classifications are stable. This stability has been examined in recent work. Simmons et al. (2008) provided intervention for students in kindergarten through 3rd grade whose scores on reading CBMs placed them at or below the 30th percentile. At-risk and out-of-risk groups were formed in beginning and end of each year based on a 30th percentile cut-point applied to CBMs and the WRMT. Results indicated little transition between at-risk and out-of-risk groups on the WRMT after 2nd grade; group membership fluctuated more frequently when reading fluency scores were used to indicate risk. Thus, measure selection impacted stability of risk designations, with the standardized measure leading to more stable classifications than the fluency measure.

In another study where latent classes were created using scores from students whose access to intervention was unknown, Catts, Compton, Tomblin, and Bridges (2012) found membership in "no-risk" or "reading disabled" classes to be relatively stable from grades 2-10

for approximately 75% of their sample, though children often changed designation within the RD classes based on word reading, comprehension, or both. When transitions between RD and no-risk classes occurred, they were often from no-risk to RD between 2nd and 4th grade, indicating emergence of "late-emerging" RD. Therefore, a second goal of the present study was to determine whether particular measure/criteria combinations, gathered after students received access to Tier II intervention in Grades 1 and 2, resulted in groups of good and poor responders that remained stable through the beginning of 3rd grade.

Intervention Responsiveness as an Indicator of Continued Risk

Given the dependency of group membership on measure/criteria combinations and the questionable stability of classifications over time, we asked whether particular RtI measure/criteria combinations could more adequately classify students, and whether those classifications would agree with RD designations made after intervention. We also questioned whether intervention provided after responsiveness classifications were made (i.e. intervention in 2nd grade) might play a role in potential misclassifications.

These issues were partially addressed in a retrospective study by Fuchs et al. (2004), where the researchers sought to identify measure/criteria combinations that, when used to form groups of good and poor 1st and 2nd grade responders to Tier II intervention, resulted in groups with significantly distinct post-intervention reading outcomes. The responsiveness of 36 first graders receiving Tier II intervention was assessed using several combinations. Five reading measures, including standardized measures and CBMs, served as post-intervention indicators of risk. Overall, combinations disagreed on good and poor responder classifications and differences on outcomes were not apparent for all RtI-derived responder groups or on all outcomes. In 1st grade, an Oral Reading Fluency (ORF)/final benchmark combination proved most stringent and

classified all Tier II students as poor responders. The WIF/median split combination resulted in groups of good and poor responders with large differences on outcomes (effect sizes > 0.90). WRMT/normalization had the next most favorable result, with large differences between groups on 4 of 5 outcome measures and with effect sizes greater than 1.0.

For 2nd graders (N=48), WIF paired with low growth identified the same students as good and poor responders as WIF paired with dual discrepancy. These methods led to more consistent and larger group differences on outcome measures compared to other methods, with average effect sizes of .85 for outcome levels and .84 for growth during intervention (Fuchs et al., 2004). Effect sizes on measures of reading comprehension were above 1.0 for each measure/criteria combination. Overall, Fuchs et al. demonstrated that measure/criteria combinations used to classify good and poor responders often disagreed. Furthermore, some measure/criteria combinations better differentiated between responder groups on post-intervention outcomes, as demonstrated by effect size differences.

One limitation of Fuchs et al.'s study (2004) is that scores on post-intervention outcomes were collected within one year of the time responsiveness was assessed, limiting the association between RtI responsiveness indicators and later reading proficiency to within-year. As a result, it is impossible to determine whether the promising early responsiveness indicators remain valid indicators of RD beyond the grade in which they were collected. Fuchs and colleagues (2008) describe a later study where they explored whether good and poor responder groups formed using various RtI measure/criteria combinations in 1st grade would generate similar group differences on outcomes at the end-of-second grade (and thus validated future indications of RD). A student was designated as RD at the end of 2nd grade if s/he scored more than 1 standard deviation below the national mean on a composite measure of sight word efficiency and on

WRMT subtests (Word Identification, Word Attack, and Passage Comprehension). First grade CBM probes for WIF were coupled with 6 RD identification methods including: final IQ-D, initial low achievement, final normalization, final benchmark, slope discrepancy, and dual discrepancy to identify the method(s) that best indicated later RD (Fuchs et al., 2008). Sensitivity (i.e. the power of a test to identify as poor responders students who later show RD) and specificity (i.e. the power of a test to identify as good responders students who continue to show typical development) requirements for RD identification were set at 0.80 each. Several 1st grade responsiveness measure/criteria combinations produced acceptable sensitivity and specificity rates for predicting RD at the end of 2nd grade. These included: (1) initial low achievement (i.e. <1SD normative sample) on WIF; (2) final normalization (standard score <90) on sight word efficiency; (3) slope discrepancy on WIF (at least 1SD below a normative sample); and (4) dual discrepancy using reading rate for level (with a score of less than 40wcpm as the cut-point) and WIF for slope (at least 1SD below normative sample).

Determining whether responsiveness groups differ on outcomes collected within-year and one year later is a good first step toward identifying RtI measures and criteria with potential for predicting distal RD. However, relying on predictions made within 1-2 years of responsiveness-group formation fails to address group instability, which primarily occurs during 1st and 2nd grade (O'Connor, Fulmer, Harty, & Bell, 2005; Simmons et al., 2008). Furthermore, RtI responsiveness indicators collected in 1st and 2nd grade and used to predict RD in these same grades may not help to identify those students who perform well on early RtI assessments, but show signs of RD later on (i.e. so called "late-emerging" poor readers; Catts et al., 2012; Compton, Fuchs, Fuchs, Elleman, & Gilbert, 2008). This failure results because late-emerging students tend to have typical reading profiles until about 3rd or 4th grade, when requirements for

reading change (Compton et al., 2008). Therefore, determining whether specific RtI measure/criteria combinations accurately predict RD in future years (especially past the end of 2nd grade) is important if responsiveness (or lack thereof) to early intervention is used for early identification of RD.

A longitudinal study conducted by Vellutino and colleagues (Vellutino, Scanlon, Zhang, & Schatschneider, 2008) provided initial evidence that students' response to intervention in 1st grade might predict reading performance in 3rd grade. The analyses and results described here constitute only the directly relevant portion of Vellutino et al.'s findings, which in whole described the reading achievement and cognitive performance of 1,373 middle-class kindergarteners. Kindergarteners who demonstrated risk on early literacy screening measures received Tier II intervention in kindergarten and were assessed on measures of word reading and comprehension through 3rd grade. At the beginning of 1st grade, the median split criterion was applied to composite scores derived from experimental measures of letter sound knowledge, decoding, and primary word identification, and from the WRMT to distinguish between responders ["No Longer at Risk" (NLAR)] and poor responders ["Continued Risk" (CR)] to kindergarten intervention. Students in the CR group received Tier III intervention in 1st grade. End-of-third grade scores on the WRMT Basic Skills Cluster (BSC) were used to further divide the CR group into those students who were "Difficult to Remediate" (DR; i.e. students with standard scores <90 on the WRMT BSC) and "Less Difficult to Remediate" (LDR; i.e. students with standard scores equal to or greater than 90 on the WRMT BSC).

Vellutino et al. found scores on word reading and reading comprehension to be distinct for the DR (lowest scores), LDR, and non-intervened (highest scores) groups at the end of 1st, 2nd, and 3rd grades. For students who showed continued risk after kindergarten intervention (*n* =

45), growth in basic skills from December to June of 1st grade accounted for 58% of the variance in end-of-third grade word reading scores; the addition of final (June) scores did not account for unique variance in outcomes beyond the growth scores. This same pattern was observed for the reading comprehension outcomes, where growth in basic skills during 1st grade intervention accounted for 36% of the variance in reading comprehension in 3rd grade and final (June) scores did not add to the prediction. Additionally, a measure of rapid letter naming accounted for 7% of unique variance in 3rd grade word reading outcomes, but accounted for no unique variance for comprehension outcomes.

Taken together, these studies suggest that a student's classification as a good or poor responder to intervention varies depending on the measures and criteria used to classify students; however, some measures (e.g. WIF) and criteria (e.g. growth, dual discrepancy, median split) may be better able to identify students who will demonstrate persistent reading difficulties despite access to quality intervention. However, few studies have attempted to measure response to instruction for a sample with access to intervention as needed during 1st and 2nd grade; results have been limited to students who received intervention from the outset of the studies.

An additional concern is sparse diversity among children in these samples. The majority of students in reviewed studies were Native English Speakers (NESs); English Learners (ELs) comprised negligible proportions of samples and whether promising RtI indicators were robust across language groups was not discussed. This issue is important due to increasing diversity of students in public schools. Although word reading develops similarly (Linklater, O'Connor, & Palardy, 2009; Mancilla-Martinez & Lesaux, 2011) and shows strong relations with fluency within NES and EL populations (Crosson & Lesaux, 2010), whether measures of word reading and text reading fluency predict distal RD for ELs within an RtI framework has yet to be

explored. Additionally, Crosson and Lesaux (2010) found that listening comprehension (and vocabulary to a lesser extent) moderated the relation between text reading fluency and reading comprehension skill for 5th grade Spanish-speaking ELs. ELs with poorly developed oral language skill performed poorly on reading comprehension measures, regardless of text reading fluency skill. So text fluency might only be able to accurately predict reading comprehension skill for ELs when measures of oral language are also considered. Therefore, it is important to investigate whether early measures of text fluency and vocabulary knowledge predict distal RD as defined by deficits in reading comprehension and vocabulary for ELs, given the complex relations between these facets of reading for this population, in particular (Crosson & Lesaux, 2010; Mancilla-Martinez & Lesaux, 2011; 2011).

The Present Study

Our study extends this literature in several ways. First, the present study is set within the context of a fully implemented RtI model (Denton, 2012), where Tier I instruction was improved through professional development and Tier II instruction was provided during 1st and 2nd grade for students at-risk. Students who showed minimal response to Tier I received Tier II intervention; poor responders to Tier II were moved to a one-on-one instructional setting where curriculum and time were reallocated to best meet individual needs. Tier III, if defined by special education placement, was controlled by school personnel; see Procedures. To our knowledge, no previous study exploring response allowed fluid transition between Tier I and II intervention groups; that is, previous studies did not include in their sample students who did not qualify for Tier II intervention when the sample was first collected (mostly between kindergarten and 1st grade), but who fell behind in later grades. A "catch and release" (Jenkins & O'Connor, 2002) process whereby access to intervention is open and receipt of intervention is contingent on

qualifying reading scores may more closely mirror ideal RtI operation in schools, and does not exclude students who have strong reading performance in early, but not later grades (i.e. potential late-emerging poor readers). Therefore, our classification procedures applied to all students, which broadens the applicability of each promising RtI indicator. We addressed the question of stability of response classifications by following students over two full years of instruction, with final assessment data gathered in the first month of 3rd grade. Additionally, we identified students as RD in the areas of word reading and reading comprehension separately. This is because a sizable proportion of students who struggle in reading after 2nd grade display deficits in only one area (Catts et al., 2012); combining performance on these skills might mask students' skill deficits in each distinct area. Also, we explored whether RtI indicators were effective in predicting RD status within each construct for ELs, specifically. Finally, we embedded immediate replication of results to address the absence of replication in earlier studies (see O'Connor & Jenkins, 1999 as an exception).

Our research questions were: (1) What combination of reading measures and criteria collected from students with access to reading intervention during 1st and 2nd grade demonstrates the most adequate sensitivity and specificity rates for predicting those children who will be identified as average or RD readers in the beginning of 3rd grade? (2) How do promising models and isolated measure/criteria combinations perform when predicting RD status of ELs? (3) Do prediction models replicate on a different cohort of students receiving the same intervention in the same schools one year later?

To address these questions, we used logistic regression and classification analyses to examine a subset of data from a longitudinal study of the impact of Tier I instruction and Tier II intervention on two cohorts of children followed from kindergarten through 4th grade (Author,

Date). The subsample of children mirrored district estimates for gender, ethnicity, language status, and achievement scores on the California Standards Test (CST).

Method

Setting

Data were collected from two Southern California school districts, henceforth referred to as District A and District B. In 2007-2008, District A served over 56,000 students. Most students (approximately 85%) identified as ethnic minorities. The largest ethnic subgroup was Hispanic (68.1% of the student body), followed by African Americans (16.3%) and Whites (10.9%). Students from other ethnicities each comprised less than 1% of the student population. Approximately 73% of students were socioeconomically disadvantaged, 43.8% of students were ELs, and 9.2% of students were enrolled in special education programs. District B served approximately 20,000 students, most of whom were ethnic minorities. Like District A, the largest ethnic subgroup were Hispanic (71.8%), followed by Whites (15.4%) and African Americans (4.4%), with other ethnicities at less than half a percent. Approximately 65% of students were socioeconomically disadvantaged, half were ELs, and 8.7% were enrolled in special education.

Students attended one of five elementary schools across Districts A and B. District A contained 3 schools and District B 2 schools. In 2007-2008, schools were comparable in size (serving approximately 450 students), except one school that served approximately twice as many students as the other schools. ELs comprised between 30 and 60 percent of students at each school; approximately 95% of ELs at each school spoke Spanish as their first language. Overall, school demographics were reflective of their district.

Participants

Data for 418 students were considered for inclusion. Cases were included in the sample if the following conditions were met:

1) Participants were 1st graders in 2007-2008 or 2008-2009,
2) Participants had *access* to Tier II intervention during 1st and 2nd grade,
3) Participants had complete or near complete data on all relevant predictors and outcomes.

Thirty one cases were excluded due to missing data on all outcomes (due to absence and unavailability for make-up testing) and/or more than half of the predictor variables. Gender distributions for removed cases mirrored that of retained cases (45% male vs. 52% male, respectively). Removed cases included proportionally more African American students compared to the retained cases (22% vs. 10%, respectively) and fewer Hispanic students (55% vs. 74%). First grade scores on the relational vocabulary subtest of the Test of Language Development (TOLD-P:3; Newcomer & Hammill, 1997) indicated equivalence between removed and retained cases; however, removed cases had lower end-of-first grade Word Identification Fluency (WIF) scores than retained cases (M = 41.6, SD = 20.9 vs. M = 50.6, SD = 23.4, respectively; p=.04).

The remaining 387 cases were split into two cohorts: students in Cohort A (N=219) attended 1st – 3rd grades from 2007-2010 and Cohort B (N=168) attended 1st – 3rd grades from 2008-2011. More than 95% received primary instruction in a general education environment. The purpose of dividing the sample was to enable exploration of the extent to which results generated from initial prediction models replicated with data from another cohort. Cohort specific participant descriptions are shown in Table 1. Overall, the demographic profiles of

students in Cohort A and B were similar to the demographic profiles of students in the schools and districts from which participants were recruited.

In 3rd grade, the sample average Peabody Picture Vocabulary Test (PPVT-R; Dunn & Dunn, 1997) standard score for Cohort B (*M*= 87.95, *SD*=10.78) was statistically equivalent to that of Cohort A (*M*=86.2, *SD*=11.44, *p*=.12; Table 1). The sample average TOLD-P:3 relational vocabulary standard score in 1st grade was approximately 1 point higher for Cohort B (*M*=8.3, *SD*=3.0) compared to Cohort A (*M*=7.3, SD=3.6, *p*=.008). According to these scores, participants from each cohort scored below the national normed average (*M* = 100, *SD* = 15) on PPVT-R in 3rd grade, and on the TOLD-P:3 (normed *M* = 10, *SD* = 3) in 1st grade. In addition, average 1st grade scores on the California English Language Development Test (CELDT), a measure of English proficiency with a range of 1 (beginning) to 5 (advanced), were equivalent between cohorts, $t(201) = -.175$, $p = .862$.

Measures

Assessments for sample selection and description. The PPVT-R (Dunn & Dunn, 1997) was used to describe receptive language in English for all students. The PPVT-R is an individually administered, norm-referenced measure of receptive vocabulary designed for individuals 2.5 years old through adult. The child selects from among four pictures, one which best represents a word read by the examiner. Standard quotient scores are reported here (raw scores standardized for age in years and months at the time of testing), with a mean of 100 and standard deviation of 15. Alternate form reliability of the standard scores range from .88-.96; test-retest reliability exceeds .9.

Subtests of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2003) and WIF (Fuchs et al., 2004) were used for intervention selection and progress monitoring. All tests were timed and individually administered.

First grade measures. First grade students received six individually administered assessments: Letter Naming Fluency (LNF), Phonemic Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF) from the DIBELS battery of assessments, WIF (Fuchs et al., 2004), and TOLD-P:3 (Newcomer & Hammill, 1997). The WIF, ORF, and relational vocabulary subtest for the TOLD-P:3 were included in analyses for this study and are described below. Remaining DIBELS subtests were used to determine intervention eligibility in 1st grade (see Intervention Eligibility). WIF was administered in the fall, winter, and spring; ORF in the winter and spring; and TOLD-P:3 in the winter.

The WIF assessment consists of word lists developed by Fuchs et al. (2004) that contain 100 isolated words randomly selected from Dolch pre-primer, primer, and first grade high frequency word lists. Students read the word list as quickly as they can. The score is the number of words read correctly within one minute, a measure of automaticity of reading skill. Alternate test form reliability exceeds .91.

DIBELS ORF measures reading rate and accuracy. DIBELS ORF passages are used nationally to identify students who need instructional support and to monitor academic progress. The student is presented with three different passages and asked to read aloud for a period of one minute for each. Scores are calculated as the number of words attempted minus errors, and the median score is used for analysis. Alternate-form reliability ranges .79 to .94 across measures, and inter-rater reliability for the first grade sample was 0.95.

TOLD-P:3 (Newcomer & Hammill, 1997) was used in 1st grade and TOLD-I:4 (Hammill & Newcomer, 2008) in 2nd grade. TOLD is an individually administered, norm-referenced measure with established reliability and validity. The relational vocabulary subtest measures a child's ability to understand and orally express the relationship between a pair of spoken words, without picture cue. Standard scores are reported here, with a mean of 10 and standard deviation of 3 for the relational vocabulary subtest. Average reliability (scorer, content sampling, and test-retest) exceeds .90 for each.

Second grade measures. The DIBELS ORF subtest, TOLD-I:4 (Hammill & Newcomer, 2008), and the Word Identification (WID), Word Attack (WA), Word Comprehension (WC), and Passage Comprehension (PC) subtests of the WRMT Normative Update (Woodcock, 1998) were administered to all 2nd grade students. ORF was administered in the fall, winter, and spring of 2nd grade, TOLD-I:4 was administered in the winter, and the WRMT subtests were administered in the fall.

The WRMT WID subtest requires students to identify words in isolation; the WA subtest requires students to apply phonic and structural analysis to pronounce pseudowords; the WC subtests require students to identify analogies to written words; and the PC subtest requires students to read 1 or 2 sentences silently with a missing word signaled by a blank space, and to supply a word that made sense in that space. Standard scores for each subtest were used in analyses unless otherwise noted (M=100; SD=15). Split half reliability ranges from .91 -.97 across subtests.

Outcome measures. ORF, PPVT-R, Test of Written Spelling (TWS-4; Larson, Hammill, & Moats, 1999), and subtests of the WRMT were used as outcome measures and were collected

in fall of 3rd grade (i.e. within the first month that students returned to school; see previous descriptions). The end-of-3rd grade CST was used as a descriptive outcome.

The TWS-4 (Larson et al., 1999) is a norm-referenced, group-administered spelling assessment. There are two alternate equivalent forms, each containing words with predictable and unpredictable spellings. Administration proceeds by dictating 20 progressively difficult words to students; the final score on the TWS-4 is the number of words spelled correctly. Test-retest and inter-rater reliability exceed .90.

The English Language Arts (ELA) section of the CST was used to further describe average and RD groups (see Procedures). Students in grades 2-11 take the CST as part of the Standardized Testing and Reporting Program; the test was designed to measure students' progress toward mastering the California state academic standards. The ELA portion of the CST for 3rd grade students contains 65 questions that assess students' skill in word analysis, reading comprehension, literary response and analysis, writing strategies, and writing conventions. Score ranges define 5 proficiency levels: Far Below Basic (FBB; 150 to 258); Below Basic (BB; 259-299); Basic (300-349); Proficient (350-401); and Advanced (402-600).

Procedures

Intervention eligibility. Participants completed universal screening in the fall, winter, and spring of each year. Those who met intervention eligibility criteria (Table 2) were referred to small group intervention. Students could be referred for intervention at any time point (fall, winter, or spring) given they met intervention criteria and had attended one of the five schools in 2007-2008. Eligible students received Tier II intervention with trained researchers and graduate students until they met pre-specified exit criteria across two consecutive time-points, at which point they were released from intervention, but continued to be monitored throughout the year.

Thus the Tier II sample received intervention as needed, with approximately 12% of the sample participating in Tier II continuously during 1st and 2nd grade. Special education services supplanted Tier II intervention if students were found eligible during the course of the study; assessment data continued to be collected, so these students' scores (n = 4 in each Cohort) are included in analyses.

Intervention selection criteria were based on DIBELS early literacy (for 1st graders) and ORF (for 2nd graders) subtests (Table 2; Kaminski & Good, 1996). These criteria differed from those recommended in the DIBELS manual due to earlier studies (Author, Date) that found published criteria too low to identify nearly all (90%) students who later demonstrated poor reading achievement. DIBELS measures were also administered every 3 weeks for progress monitoring and scores were used to identify students whose performance warranted exit from intervention. Exit criteria (Table 2) closely mirrored the "low risk" cut-offs from the DIBELS manual.

Tier II intervention. The Tier II intervention consisted of small group (two or three students) instruction for 25-30 minutes in 1st and 2nd grade, four times per week (up to 2 hours per week). Intervention was offered September through April during the academic year. Intervention for 1st graders and for 2nd graders whose early reading scores suggested need for instruction in decoding and word reading was based on *Sound Partners* (Vadasy et al., 2005) and included letter-sound correspondence, decoding, sight word identification, and reading of sentences and decodable books. These activities have generated significant improvement for low-skilled 1st and 2nd grade students (Vadasy, Jenkins, Antil, Wayne, & O'Connor, 1997; Vadasy, Sanders, & Tudor, 2007). Second grade students whose decoding and fluency scores indicated need for practice reading connected text rather than phonics and word/sentence reading

participated in a more advanced Tier II instruction, which included word study (with multisyllabic words), vocabulary, and comprehension activities, reading and rereading books at students' current reading level, and brief spelling and sentence-writing opportunities. This multi-component researcher-created intervention incorporated strategy instruction for teaching word reading (e.g. the BEST strategy, O'Connor, 2007), vocabulary acquisition (Beck, McKeown, & Kucan, 2002), and comprehension (see Klingner et al., 2007); repeated and themed reading were also incorporated for fluency practice (Hudson, 2012).

ELs were integrated with NESs in intervention groups; that is, ELs did not receive specialized EL instruction within the Tier II intervention apart from what was provided to all students. Although ELs typically have different instructional needs than NESs, especially in the areas of oral language and comprehension (e.g more focus on developing oral language skill in the early grades; greater incorporation of gestures and pictures and greater emphasis on common, but unknown words, when teaching vocabulary; Linan-Thompson & Vaughn, 2003), the oral language skill of NESs in our study was quite low on average; their oral language and vocabulary needs mirrored those of the ELs (see results). Small group sizes (1 to 3 students) allowed for increased opportunities to practice speaking in English for all tutored students. In addition, illustrations were used to elucidate word and text meaning as appropriate, and graphic organizers, semantic maps, questioning strategies, and teacher modeling of appropriate responses were used to facilitate reading comprehension for all tutored students (Linan-Thompson & Vaughn, 2003; Klingner et al., 2007).

Intervention students were regrouped monthly based on their individual needs and assessment scores. Students who demonstrated little progress (based on progress monitoring every three weeks) were moved to a one-on-one setting where interventionists administered

individualized intervention (similar to a Tier III setting). Poor response to Tier II instruction was mainly attributed to a mismatch between student and instruction (e.g. the instruction was not intensive or individualized enough) and not to poor fidelity, since tutor fidelity was high (see below). The intervention curriculum shifted between the researcher-created intervention and *Sound Partners* as needed, and time allotments for particular activities (e.g. word reading, fluency practice, phonemic awareness) were reorganized to support student learning. Special education referrals were initiated by classroom teachers and other stakeholders that did not include the research team, and referrals and placement in special education cannot be directly tied to failure to progress during the intervention. However, the four students in our sample who were identified for and received services in special education during the project had also demonstrated minimal progress during intervention, and had received 1 on 1, targeted intervention support after failure to progress in response to small group Tier II instruction.

Intervention was delivered by project staff during the regular school day. Students were pulled from their general education classroom at times that did not conflict with ELA instruction (see "Tier I Instruction") or EL programming, which varied from 30-150 minutes per week depending on student level of English proficiency.

Tutors and training. Tutors included experienced special education teachers, classroom teachers, graduate students, teacher credential candidates, and teacher assistants. Across the staff, 61% of these individuals were tutors for the entire three years of this research; 88% were with the project at least two years. All tutors received training from the PI in instructional delivery of the specific curricula for each grade level; lead tutors at each school received an additional 30 hours of training. The initial four-hour training for all tutors included a theoretical introduction of each reading activity, modeling of the activity, guided practice, and independent practice in

small groups with observation, feedback, and discussion of common problems. Tutors received a teacher manual generated by the PI and Co-PI, which besides the student curricula and teacher scripts, included a pacing guide for daily lessons, a pacing guide for monthly progress based on average progress, and flow charts linking specific types and levels of activities to progress monitoring benchmarks. This initial training was supplemented by bi-monthly follow up training where new activities were introduced, common issues noted during field observations were discussed, and additional practice provided.

Tier II treatment fidelity. An experienced general or special education teacher was designated as the lead tutor at each site. The lead tutor informally observed and provided feedback to tutors daily; s/he oversaw weekly progress of students and modifications to the monthly lesson plans, and collected daily activity logs completed by tutors for each of the small groups. The lead tutors also formally observed lesson implementation once per month for two lessons for each of the tutors at each site. Researcher-created observation forms were used to track treatment fidelity. Reviewers looked for two indicators of treatment fidelity when conducting observations and reviewing activity logs: completion of the each of the steps/activities outlined in the teacher scripts and student growth in DIBELS measures. Poor rate of growth triggered a conference where activities, group structure, and/or pacing were changed. Fidelity was computed as a ratio of all observed actions to all actions expected. If any observation fell below 85% fidelity, the tutor was provided coaching and feedback, followed by co-teaching with the lead tutor until acceptable fidelity was reached. This occurred no more than 6 times per year, with an immediate rise in fidelity to acceptable levels (i.e. at least 85%) after corrective action was applied. The average fidelity rating for *Sound Partners* was 92.06% and for second grade curricula 89.4%. Across curricula, fidelity ratings ranged from 83% - 100%.

Tier I instruction. Teachers in all grades used the Houghton Mifflin Reading California Language Arts Curriculum (Cooper et al., 2003), which is aligned with California State Standards for reading and writing development. The curriculum encourages development of oral language, phonemic awareness, letter recognition, phonics and blending skills, and high frequency vocabulary recognition in the early grades. In later grades, increasing emphasis is placed on comprehension strategies, including: prediction/inference, monitoring/clarifying, questioning, summarizing, and evaluating text. Instruction in spelling and writing is also incorporated. Theme-based stories form the basis for published teacher lessons, which include elements such as activating background knowledge, pre-teaching vocabulary, options for teacher modeling of decoding and comprehension strategies, and graphic organizers for comprehension.

Teachers at each of the 5 schools participated in 120 hours of language arts professional development in reading as part as two California mandates: Assembly Bill 1485 (AB1465) The Reading First California Technical Assistance Center and Assembly Bill 466 (AB466) for under-performing districts. Included in training were strategies for direct instruction, guided reading, frequent diagnostic assessments, and providing for universal access (small group instruction and differentiated instruction); a pacing guide for coverage of the California Language Arts standards was also included. Principals at these schools were required to incorporate these elements in teacher evaluations and classroom observations.

Definition of reading disability. We formed average and RD groups based on fall-of-3rd grade reading performance in word reading and comprehension separately. Vellutino et al. (2008) noted that outcomes collected after the summer between grades may be more accurate reflections of proficiency or risk since students are required to maintain gains over summer with little to no instruction (see also O'Connor et al., 2005; Vellutino et al., 1996). Based on this

"summer slippage," we reasoned that data from the first month of 3rd grade would yield more valid RD classifications than those collected immediately following intervention. We then used scores on 1st and 2nd grade measures to predict later group membership for students who had access to intervention during 1st and 2nd grade. Versions of this approach have been used by Catts et al. (2012) and Vellutino et al. (2008). This method allows control over the definition of "average and above (or poor) reading" in later grades and also preserves the ability to use continuous measures in earlier grades to predict a student's group membership. Furthermore, by separating prediction of RD into word reading and comprehension constructs (see below), we alleviated risk of classifying students with strengths in one area and deficits in the other as "average" readers overall.

We had access to 3rd grade measures of word reading, decoding, fluency, spelling, vocabulary, word comprehension, and passage comprehension for all students. Theory and empirical research guided our placement of measures of word reading, decoding, fluency, and spelling on one construct, and measures of vocabulary, word comprehension, and passage comprehension on a separate construct (Cain & Oakhill, 2011; Catts et al., 2012; Chall, Jacobs, & Baldwin, 1990; Ehri, 1998; Hart & Risley, 1995; Hudson, Torgesen, Lane, & Turner, 2012; Morris et al., 2012; Schwanenflugel, Meisinger, Wisenbaker, Kuhn, Strauss, & Morris, 2006; Simmons et al., 2010; Vellutino et al., 2008; Verhoeven, van Leeuwe, & Vermeer, 2011; Wieser & Mathes, 2011).

To verify whether skills mirrored the theorized and/or empirically suggested constructs for students in our sample, we conducted a Confirmatory Factor Analysis (CFA) with two factors representing the 7 measures (Table 3). We placed the measures of word reading, decoding, fluency, and spelling on a Word Reading-Fluency (WR-F) factor and measures of vocabulary,

word comprehension, and passage comprehension on a Comprehension-Vocabulary (C-V) factor. For Cohort A, fit statistics indicated adequate model fit, $X^2(13) = 26.689$, $p = .013$; CFI = .984; TLI = .974; RMSEA = .069, CI = .03 to .11 (Raykov & Marcoulides, 2006). The loadings of measures on each factor were moderate to large and significant, and factors were correlated ($r = .837$, $p < .001$) as expected. We recognized that the loading for the PPVT on the C-V factor was low compared to C-V measures; constraining the factor loading to zero with data from Cohort A resulted in no change in the correlation between factors and a significant increase in Chi-Square ($\Delta X^2 = 26.54, df = 1; p < .001$), and therefore significant deterioration of model fit. CFA with Cohort B yielded similar patterns with good model fit (Table 3). Therefore, theoretical and statistical results supported our two factor solution to describe skill in WR-F and C-V.[i] **Creating RD and average reader groups.** The procedure used to form WR-F composites was modeled after that used by Catts et al. (2012): each of the 3rd grade fall ORF, WRMT WID and WA, and TWS-4 raw scores were converted into *Z* scores using sample-based means and standard deviations. The *Z* scores for each measure were combined to form a composite score for WR-F outcomes. Finally, the composite WR-F score was converted into a sample-based *Z* score. The resulting variable reflected a student's WR-F performance in the fall of 3rd grade in standard deviation units relative to same-grade peers attending similar schools and receiving the same access to equivalent Tier II intervention. The steps used to create average and RD groups under the WR-F construct and C-V construct (see below) were repeated for students in Cohort B.

The procedure used to create the C-V composite variable was identical to that used to create the WR-F variable except for the selection of measures. Fall of 3rd grade raw scores for WRMT WC and PC and for the PPVT were used to create the C-V composite. The resulting

variable reflected a student's C-V performance in fall of 3rd grade in standard deviation units relative same-grade peers attending similar schools and receiving the same access to equivalent Tier II intervention.

Composite scores provided a holistic description of reading performance within each construct and eliminated error associated with using single measures to identify RD. We recognized that an alternative approach to creating composites might have been to use the factor scores resulting from the CFA within each cohort. We chose not to utilize factor scores for a few reasons. First, we thought it important that measures within each construct contribute equally to composite scores used in RD classification. That is, we wanted to avoid misclassification of students with relative strengths or weaknesses in particular skills (e.g. untimed word reading) that may result from overemphasizing the contribution of those skills to the construct. Furthermore, the CFA was used simply to test factor structure across cohorts and not to determine relative strength in factor loadings. Had we utilized factor scores in creating composites, the interpretation of WR-F and C-V composites would have been slightly different across cohorts due to small differences in factor loadings. For example, the C-V composite would have relied more on the WRMT PC scores for Cohort A and on WRMT WC scores for Cohort B. Overall, from a statistical viewpoint, using weighted factor scores may appear preferable; however, from a theoretical viewpoint, we wanted to ensure that students were not misidentified as "average" or "RD" readers based on scores for a single and perhaps overly-influential measure. In addition, we thought it important that interpretations of performance on WR-F and C-V remained consistent across cohorts to facilitate comparison; creating equally-weighted composites accomplished these aims. The standard error for the composite scores (*SE* = .07) was used to polarize RD and average reader groups; average readers scored equal to

or better than .93 standard deviations below the local-normative mean, and RD students scored ≤1.0 standard deviations below the mean. We reasoned that using local norms (as opposed to national norms) to form RD and average reader groups was appropriate given our intent to predict RD status for students experiencing similar Tier I conditions and having access to similar Tier II interventions – in this way we could assert that compared to average readers, RD readers were responding differently to nearly identical instructional opportunities, and therefore may have had different instructional needs. Also, although national norms have become more representative to ethnic minorities and students living in poverty, score distributions based on national norms for some assessments may not have been appropriate indicators of RD for our low income, ethnically and linguistically diverse sample[ii]. We sought to weigh students' responsiveness to intervention against that of their demographically similar peers.

Tables 4 and 5 show RD prevalence estimates, and means and standard deviations for 3rd grade outcomes and an external criterion (i.e. the CST) by RD and average groups within the WR-F and C-V constructs for each cohort, respectively. Table 6 displays prevalence of WR-F, C-V, and combined disabilities within each cohort, as well as minutes of intervention received by RD and average reader groups. As can be seen in Tables 4 and 5, significant and substantive differences and moderate effect sizes on reading outcomes between RD and average reader groups supported our group classification procedure. Additionally, differences between groups on 3rd grade CSTs were significant (*p*s <.001) and substantively meaningful under the WR-F and C-V constructs, with effect sizes of .39 and .49 respectively in Cohort A, and .48 for each construct in Cohort B. In each case, average readers scored in the Basic range (which was also the average score range within the districts) and RD readers scored in the Below Basic/ Far Below Basic range. It is important to recognize that some students in each cohort displayed

combined deficits (i.e. deficits in WR-F and C-V; see Table 6). That students with dual-deficits were represented within each construct might explain the significant differences between average and RD reader groups for some cross-construct measures (e.g. for WRMT PC under the WR-F construct; Tables 4 and 5).

We defined late-emerging poor readers (LEPRs) as students who did not meet intervention criteria in 1st or 2nd grade and who did not meet our RD classification criteria based on scores for 2nd grade measures, but who met the criteria based on 3rd grade scores. Though we did not utilize latent class or transition models to identify LEPRs (Catts et al., 2012; Compton et al., 2008), we felt our process was adequate for the purpose of determining whether promising responsiveness indicators were able to identify these children by 3rd grade, since they were indeed students with strong reading performance in early, but not later grades.

Results

We used logistic regression with data from Cohort A to identify the 1st and 2nd grade measures that best predicted RD status. Next, we paired significant predictors with various RtI criteria to determine whether isolated 1st or 2nd grade measure/criterion combinations adequately predicted average and RD status in 3rd grade. Sensitivity and specificity of measure/criteria combinations were evaluated against each other and the field standard of .9 for sensitivity and .8 for specificity (Jenkins et al., 2003). We also explored correct classification of ELs and adjusted specificity rates for promising RtI measures collected before the end-of-second grade. The adjusted specificity rates describe the power of RtI measures to identify children who will be average readers in 3rd grade *as long as they have access to intervention as needed.* We chose not to use Receiver-Operating-Curve (ROC) analysis as our primary investigation tool; rather, we selected RtI criteria that schools might be better able to utilize, such as benchmarks and

percentile scores, to determine risk. Nevertheless, AUC and specificity values with associated sensitivity of at least .90 are provided for promising measure/criteria combinations in Table 11 to enable comparisons across studies. Finally analyses were repeated with Cohort B to explore replication.

Data Screening and Preliminary Analyses

Due to the multilevel nature of our data, we sought to determine whether significant variance in WR-F and C-V outcomes was present between classrooms. We estimated a two-level unconditional model with students at Level 1 and classrooms at Level 2 for each outcome within each cohort using the HLM 7 software (Raudenbush, Byrk, Cheong, Congdon, & Toit, 2011). Intraclass Correlation (ICC) was calculated to identify the proportion of variance in outcomes that existed between classrooms (Raudenbush & Byrk, 2002). The outcomes used for these analyses were scores on each of the WR-F and C-V composites. Within Cohort A, the ICC indicated that 2% of variance in WR-F composite scores was between classrooms, χ^2 (20) = 29.25, p = .08, and 3% of variance in C-V composite scores was between classrooms, χ^2(20) = 35.57, p = .02. Although significant, this proportion of variance was negligible. Identical models were repeated to determine whether significant variance in WR-F and C-V outcomes was present between classrooms for students in Cohort B. The ICC indicated that 3.4% of variance in WR-F composite scores was between classrooms, χ^2 (18) = 25.26, p = .147, and 1.8% of variance in C-V composite scores was between classrooms, χ^2 (18) = 20.29, p = .316. In sum, variance between classrooms was minute for each model estimated, with no variance exceeding 3.5%. Additionally, with one exception (i.e. the C-V model for Cohort A), the chi-square test of the null that variance at Level 2 is equal to zero was non-significant for all models tested. Therefore, data were analyzed without estimating classroom effects on outcomes.

Means and standard deviations are provided for demographic and assessment variables for each Cohort in Table 1. Non-parametric tests of distributional differences in gender, ethnicity, and EL status between cohorts revealed no significant differences, all *p*s> .12.

Table 1 also provides means and standard deviations for 3rd grade outcomes and for 1st and 2nd grade predictors for each cohort. *P*-values associated with univariate test of differences between cohorts are provided where significant. The significance level was adjusted to $\alpha = .003$ to correct for the 16 comparisons between groups. Significant differences were evident for the 2nd grade WA and PC subtests for WRMT, where scores for Cohort B were higher on average than those of Cohort A.

Correlations among 3rd grade outcomes are provided for each cohort in Table 7; correlations among predictors and between predictors and outcomes are provided for each cohort in Table 8. Note that in Tables 7 and 8, correlations for Cohort A are reflected on the rows and below the diagonal and those for Cohort B are on the columns and above the diagonal. Correlations among WR-F measures and C-V measures were moderate to strong and significant. Overall, correlations for Cohort B were stronger than those for Cohort A.

To determine if multicollinearity posed a serious issue, we calculated the Variance Inflation Factor (VIF) for each highly correlated variable. VIFs were compared to a maximum acceptable value of 10 (Cohen, Cohen, West, & Aiken, 2003); no VIFs met or exceeded this value. Therefore, multicollinearity was not considered to significantly affect model results.

Cohort A

A MANOVA was executed in SPSS 18 to determine whether scores on outcome composites and predictor variables varied by gender, ethnicity, or EL status. Box's *M* test warranted acceptance of the assumption of homogeneity of covariance matrices at $\alpha=.01$, $F =$

1.176, $p > .01$. Applying a more conservative significance level is suggested when interpreting results of Box's *M* due its notorious sensitivity to non-normality (Raykov & Marcoulides, 2008). Levene's test of equal error variances for univariate analyses revealed no significant differences and therefore no univariate homogeneity assumption violations.

The MANOVA of all variables by gender was significant, Wilks's $\Lambda = .861$, $p = .001$. Univariate tests revealed significant differences on the composite score for WR-F, $F(1, 198) = 3.79$, $p = .05$, and V-C, $F(1,198) = 5.68$, $p = .013$. In each case, males outscored females. Therefore, gender was entered as a covariate for logistic regression analyses.

The MANOVA of all variables by EL status, Wilks's $\Lambda = .953$, $p = .496$, and ethnicity (Wilks's Λ Hispanic = .923, $p = .113$ and Wilks's Λ African American = .960, $p = .641$) were not significant. Univariate tests revealed no significant differences on any variables at a significance level of .05, though there was a trend of lower scores for Hispanic students on the C-V composite, $F(1,198) = 15.89$, $p = .07$.

Cohort B

Similar to Cohort A, MANOVAs were executed with Cohort B to determine if scores on predictors or outcomes varied by student characteristics. Box's *M* test was non-significant at $\alpha = .01$ ($p = .02$) indicating no violation of the homogeneity of variance/covariance assumption. A MANOVA of all variables by gender was non-significant, Wilks's $\Lambda = .947$, $p = .720$; however, univariate tests revealed significant differences on 1st grade WIF, $F(1, 151) = 4.41$, $p = .04$, and 1st and 2nd grade ORF, $F(1,151) = 5.8$, $p = .02$ and $F(1,151) = 4.9$, $p = .03$, respectively. Unlike Cohort A, females outscored males.

A MANOVA of all variables by EL status was significant, Wilks's Λ = .860, p = .025. Univariate tests revealed no significant differences on any measure at α = .05, and trend favoring native English speakers on the composite variable for C-V at α = .10, $F(1,151) = 3.74$, p = .06.

A MANOVA of all variables by ethnicity was significant for Hispanic (Wilks's Λ = .851, p =.02) and non-significant for African American (Wilks's Λ = .926, p = .428). Univariate tests for Hispanic revealed lower scores on 1st grade TOLD-P:3, $F(1,151) = 6.6$, p = .01, and ORF, $F(1,151) = 4.53$, p = .04; 2nd grade ORF, $F(1,151) = 4.0$, p = .05, WID, $F(1,151) = 4.64$, p = .03, and PC, $F(1,151) = 5.6$, p = .02. Univariate tests for African American revealed no significant differences on any measure.

Overall, MANOVA results were similar across cohorts. The exception was the gender effect. Males outscored females on 3rd grade WR-F and C-V composite measures in Cohort A, and females outscored males on select 1st and 2nd grade predictor variables in Cohort B. Although significant at α = .05, no gender differences were significant after significance levels were adjusted for multiple comparisons using Bonferroni correction (i.e. at α = .004). Nevertheless, the gender coefficient was significant in WR-F and C-V regression analyses with Cohort A (see below) and was therefore a useful covariate.

Logistic Regression

Word reading/fluency. We used logistic regression with data from Cohort A to identify 1st and 2nd grade reading measures that best predicted reading proficiency/disability in 3rd grade. Recall that students were identified as RD if their scores on the composite WR-F variable (comprising 3rd grade ORF, WID, WA, and spelling measures) fell 1.0 standard deviations or more below the sample mean. Predictors were entered stepwise, with spring-of-1st grade

measures in the first step and if significant, included in Step 2 with fall, winter, and spring-of-2nd grade measures. Demographic covariates were entered in the first step and retained if significant.

Logistic regression coefficients, standard errors, *p*-values, and odds ratios are provided in Table 9. The final model included only significant predictors from Step 2, and is displayed below. Model sensitivity and specificity were 88.9% and 86.2%, respectively. Three cases (2 NES, 1 EL; 2 in WR-F only, 1 in WR-F + C-V) with RD and 25 average readers (10 NES, 15 ELLs) were misclassified.

$$RD = 26.167 - 2.302*Gender - 0.101*ORF\ 2^{nd} - 0.209*PC\ 2^{nd}$$

Note that beta coefficients for each predictor express the change in log odds of being in the group coded 1 (i.e. the RD group) for every 1 unit change in the predictor. All significant coefficients were in the expected direction; that is, for each measure, unit increases in the predictors were associated with reduced log odds that a case would belong to the RD group. Exponentiation of log odds eases interpretation; subtracting 1 from the odds and multiplying by 100 provides the percent increase (for positive values) or decrease (for negative values) in odds of being in the RD group for every 1 unit increase on the predictor variable. For example from the final model illustrated in Table 9, 2nd grade ORF had an odds ratio (or simply odds) of .904. Subtracting 1 from this value and multiplying by 100 shows that for every 1 unit increase on 2nd grade ORF, the odds of being in the RD group decreased by approximately ten percent[iii].

The purpose of the logistic regression analysis was to identify 1st and 2nd grade measures best able to distinguish between RD and average reader groups, so tests of model deterioration or improvement were not of primary relevance. Nevertheless, AIC and BIC values were lower at each step than the Null model, indicating superior model fit for each of the conditional models compared to the Null (Raykov & Marcoulides, 2008). Traditional Chi-Square statistics are not

provided in Mplus. Therefore, Chi-square difference tests of each step compared to the Null model were conducted using changes in negative loglikelihood values with degrees of freedom equal to the number of independent variables in the model[iv]. The change in negative loglikelihood of the Null compared to Step 1 (-loglikelihood null = 5489.4, *df* = 0; Step 1 = 1826.37, *df* = 4) was significant, indicating significant improvement in the model with the addition of predictors, $\chi^2(4) = 3663.03$, $p < .001$. The change in negative loglikelihood from the Null to Step 2 was also significant, $\chi^2(6) = 1407$, $p < .001$, as was the final model from each of the Null, $\chi^2(3) = 3702.34$, $p < .001$, Step 1, $\chi^2(1) = 39.31$, $p < .001$, and Step 2, $\chi^2(3) = 2232.23$, $p < .001$.

Comprehension/vocabulary. As with word WR-F outcomes, logistic regression was used to identify 1st and 2nd grade reading measures that were most useful in the prediction of RD in C-Vin 3rd grade. Model building procedures mirrored those of the WR-F models.

Logistic regression coefficients, standard errors, *p*-values, and odds ratios are provided in Table 10. The final model is displayed below.

$$RD = 4.576 - 0.923*Gender - 0.038*ORF\ 2^{nd} - 0.408*TOLD\ 2^{nd}$$

Model sensitivity and specificity for Cohort A were 84.4% and 76.7% respectively. Five cases with RD (2 NES, 3 ELs; all with RD in C-V only) and 41 average readers (15 NES and 26 ELs) were misclassified. As with the WR-F model, significant coefficients were in the expected direction. Additionally, AIC and BIC values favored the conditional models over the Null and the final model over Steps 1 and 2. Chi-square change statistics indicated significant model improvement at each step and the final model compared to the Null.

Replication. Our second research question queried the extent to which logistic regression models and classification analyses (see below) executed with Cohort A replicated with a new

sample. For analyses on WR-F outcomes, prior probabilities were set to the proportion of cases with RD in WR-F from Cohort A (13%). Setting prior probabilities to the proportion of cases with RD in WR-F from Cohort B (i.e. 14%) resulted in no change in model sensitivity and specificity indices. Therefore, cases with predicted probabilities at or above 13% were classified as RD for Cohorts A and B. The same method for setting prior probabilities was adopted for analysis of C-V outcomes. In this case, priors were set to 14% for each cohort. Previous studies with similar analyses have also used proportions associated with RD groups as the prior probability cut-off (e.g. Compton, D. Fuchs, L. S. Fuchs, & Bryant, 2006); this procedure appears useful for making cut-point decisions for priors in logistic regression.

Recall that the final logistic regression model used for the prediction of RD in WR-F for Cohort A was $RD = 26.167 - 2.302*Gender - 0.101*ORF\ 2^{nd} - 0.209*PC\ 2^{nd}$, with sensitivity of 88.9% and specificity of 86.2%. Replication of the final model on Cohort B resulted in sensitivity of 80% and specificity of 89.2%. Five cases with RD (1 NES, 4 ELs; 3 in WR-F only, 2 in WR-F + C-V) and 15 average readers (10 NES, 5 ELs) were misclassified.

The final logistic regression model for the prediction of RD in C-V for Cohort A was $RD = 4.576 - 0.923*Gender - 0.038*ORF\ 2^{nd} - 0.408*TOLD\ 2^{nd}$, with sensitivity of 84.4% and specificity of 76.7%. Replication of the model on Cohort B resulted in sensitivity of 82.1% and specificity of 69.8%. Five cases with RD (all ELs; 4 in C-V only, 1 in C-V + WR-F) and 42 average readers (19 NES, 23 ELs) were misclassified.

Classification Analyses

We created two-by-two contingency tables posing responder status against RD classification to determine whether any significant predictor of RD, when paired with some criterion, adequately identified average and RD readers. Significant predictors of RD in WR-F

were 1st grade WIF (spring), 2nd grade ORF (spring), and WRMT PC (fall). Although WIF was no longer significant once 2nd grade measures were added to the model, WIF was included in the classification analysis to explore classification rates for RD resulting from a 1st grade reading measure. To note, we chose to explore end-of-1st grade WIF as an isolated predictor rather than end-of-1st grade ORF because WIF and not ORF emerged as a significant predictor of RD when both measures were included in the model together (see Table 9). Sensitivity and specificity of all explored measure/criteria combinations are provided in Table 11; specificity estimates were adjusted on 1st grade and fall-of-2nd grade measures with promising sensitivity to account for the possibility that intervention received in 2nd grade influenced 3rd grade classification. Misclassified RD readers met RD criteria in 3rd grade, but were classified as good responders. Misclassified (or over-classified) average readers did not meet RD criteria in 3rd grade, but were flagged as poor responders.

Since no published benchmarks exist for WIF, 1st grade WIF was paired with 25th and 33rd percentile criteria. Children scoring below the relevant percentile cut-point for this and other measures using percentile criteria were designated poor responders. Under WIF/25th percentile, 75% of children with RD and 81% of average readers were correctly classified. WIF/33rd percentile correctly classified 82.1% of RD readers and 76.1% of average readers. Under WIF/33rd, all 5 misclassified RD readers were ELLs; 2 received 1st grade intervention and 3 didn't qualify for intervention until 3rd grade. Of the 44 over-classified readers, 20 were NESs and 24 were ELs. Furthermore, 13 (29.5% of those over-classified) never received intervention; the remaining 31 students (70.4% of those over-classified) received intervention in 1st & 2nd grade (n=6) or 2nd grade only (n= 25). Thus, the adjusted specificity rate: 93%[v]. Lastly, each of

the WIF combinations identified children with RD in WR-F only and with combined RD (RD in WR-F + C-V).

Second grade ORF was paired with final benchmark, low growth, dual discrepancy, and percentile cut criteria. According to DIBELS benchmarks (Good & Kame'enui, 2001), 90 wcpm by spring-of-2nd grade specifies the no-risk cut-off; therefore, children reading 90 or more wcpm were categorized as good responders and those reading fewer than 90 wcpm were categorized as poor responders. Under ORF/final benchmark, 100% of children with RD and 55.9% of average children were correctly classified. To improve specificity, the final benchmark criterion was reduced to the highest score in the RD group: 75wcpm. Under the new criterion, sensitivity remained at 100%[vi] and specificity improved to 78.8%.

The ORF/low growth criterion was specified such that children with growth ≤ 1 standard deviation below the sample average growth (*M*=33.8 wcpm, *SD*=6.3 wcpm) were classified as poor responders. This criterion produced poor sensitivity rates and adequate specificity, correctly classifying 37.9% and 89.4% of RD and average readers, respectively. Children were identified as poor responders under ORF/DD if they met both ORF/final benchmark and ORF/low growth criteria. Under ORF/DD, 40.7% of RD and 95.1% of average readers were correctly classified.

The cut-score for the 25th percentile on ORF was 61 wcpm. Under this criterion, sensitivity and specificity were 85.2% and 83.2% respectively. Six RD students (1 NES, 5 ELLs) and 29 average students (12 NES, 17 ELs) were misclassified. Under ORF/33rd percentile (ORF < 67 wcpm), sensitivity improved to 93% with specificity of 75.5%. ORF/33rd percentile correctly classified an additional 4 RD readers and misclassified an additional 14 readers compared to ORF/25th percentile.

Finally, 2nd grade WRMT PC was paired with 25th and 33rd percentile criteria. Sensitivity and specificity for PC/25th were 79.3% and 81.6%, respectively. Six RD students (3 NES, 3 ELs) and 34 average readers (12 NES, 22 ELs) were misclassified. Of the 6 misclassified RD readers, 4 did not receive intervention until 2nd grade or later. Of the 34 over-classified average readers, 14 (41.1%) received intervention in 2nd grade. The remaining 20 (58.8%) did not, resulting in an adjusted specificity of 89.2%. The 33rd percentile criterion improved sensitivity to approximately 90%, with specificity of 73.5%. Two of the 3 misclassified RD readers were NESs, and none received intervention until 2nd grade or later. Of the 49 over-classified readers (18 NES, 31 ELs), 16 (32.6%) received 2nd grade intervention and 33 (67.3%) did not, resulting in an adjusted specificity of 82.2%.

Significant predictors for RD in C-V were 2nd grade TOLD-I:4 (winter) and 2nd grade ORF (spring). Although 1st grade TOLD-P:3 (winter) was no longer significant after the inclusion of 2nd grade measures, it was retained in classification analyses to determine whether it was useful as an isolated predictor. To identify good and poor responders, 1st and 2nd grade TOLD were paired with 25th and 33rd percentile criteria and ORF was paired with final benchmark, low growth, DD, and percentile-cut criteria.

First and 2nd grade TOLD measure/criteria combinations yielded poor sensitivity and specificity indices overall. In each case, classification rates fell far below field standards. ORF/criterion combinations were also associated with poor classification rates overall, with no combination approaching acceptable classification rates for RD or average students (Table 12). Adjusted classification rates were not explored due to poor initial classification.

Replication. We replicated all classification analyses using data from Cohort B. Like Cohort A, WIF/33rd percentile yielded adequate sensitivity (88.5%) and specificity (76.1%) when

predicting RD in WR-F; all misclassified RD readers and 14 of 33 over-classified average readers were ELs. Additionally, 84.8% of over-classified average readers received intervention in 2nd grade, resulting in adjusted specificity of approximately 96%. The 25th percentile criterion for 1st grade WIF performed slightly better with Cohort B scores compared to Cohort A, with sensitivity of 80.8% and specificity of 81.9%. Three of 5 RD readers and 10 of 25 average readers that were misclassified were ELs. All but one over-classified readers received intervention in 2nd grade, resulting in adjusted specificity of more than 99%.

The final benchmark, low growth, dual discrepancy, and percentile-cut criteria used to identify good and poor responders according to 2nd grade ORF were identical to those used with Cohort A, and results were comparable. Under final benchmark, sensitivity was 100%. Although no RD readers were misclassified, more than half of the average readers were, with specificity of 43.2%. To improve specificity, the final benchmark criterion was lowered to the highest score earned by any RD reader on ORF (i.e. 70 wcpm). This change did not impact sensitivity by design, and improved specificity to 69.1%.

Similar to results from Cohort A, the low growth criterion resulted in poor sensitivity and excellent specificity. Sensitivity and specificity were 69.6% and 95%, respectively. Case classification for DD was identical to that for low growth. Every case with growth less than or equal to 1 standard deviation below the sample mean also had spring ORF scores lower than 90 wcpm. Thus, sensitivity and specificity were also 69.6% and 95%, respectively.

ORF/25th percentile produced a cut-score of 61 wcpm. Under this criterion, 84.6% of RD readers and 85.6% of average readers were correctly classified; these rates are nearly identical to those from Cohort A. Three of the 4 misclassified RD students and 7 of the 18 over-classified students were ELs. The cut-score produced by the 33rd percentile criterion was 67 wcpm. This

criterion correctly classified an additional 3 RD readers, with a sensitivity of 96.2%. Nine additional average readers were misclassified as RD, yielding a specificity of 79.1%. Again, sensitivity and specificity here mirrored results from Cohort A.

The 25th percentile criterion for 2nd grade PC provided the best balance between sensitivity and specificity. Five RD (2 NES, 3 EL) and 23 average readers (11 NES, 12 ELs) were misclassified, resulting in sensitivity of 80% and specificity of 83.6%. Of the over-classified average readers, 56% received intervention in 2nd grade, resulting in adjusted specificity of 93%. An additional 2 RD readers were correctly classified under the 33rd percentile criterion, improving sensitivity to 88%. Thirteen additional average readers were misclassified, 9 of whom received 2nd grade intervention. Thus, the adjusted specificity: 90%. Again, these results are similar to those of Cohort A.

Significant predictors of RD in C-V were 1st and 2nd grade TOLD and 2nd grade ORF. As found with Cohort A, classification rates were poor for each measure/criteria combination explored (Table 12). An exception was the ORF/33rd percentile combination, which yielded relatively reasonable sensitivity and specificity of 78.6% and 76.3%, respectively.

Late Emerging Poor Readers (LEPRs)

We identified 9 LEPRs (4 in WR-F, 4 in C-V, and 1 in both) in Cohort A and 4 LEPRs (1 in WR-F, 3 in C-V) in Cohort B, which corresponds to sample prevalence rates of 4% and 2.3% across cohorts. We explored whether the most promising measure/criteria combinations for identifying RD were able to correctly classify LEPRs. The WIF/33rd, PC/25th, and ORF/25th correctly classified between 60 and 100% of LEPRs in WR-F across cohorts (Table 11). Due to the low prediction power, and therefore low utility, of measure/criteria combinations for RD in C-V, we did not explore LEPR predictions in C-V.

Discussion

We examined the ability of 1st and 2nd grade reading measures to predict RD in 3rd grade for students who received *access* as needed to Tier II reading intervention during 1st and 2nd grade. We paired responsiveness criteria (e.g. final benchmark, dual discrepancy, percentile-cuts) with predictive 1st and 2nd grade measures to identify measure/criteria combinations most effective in identifying children classified as RD by the beginning of 3rd grade. We investigated classification accuracy for ELs and also explored the extent to which results replicated with a second cohort of students (i.e. Cohort B) whose instructional access (including schools, teachers, and interventions) was nearly identical to the initial cohort (i.e. Cohort A), but who attended 1st grade the year after Cohort A.

We extended prior work on the use of responsiveness measures to identify reading risk or RD (e.g. Compton et al., 2006; Compton et al., 2012; Fuchs et al., 2004; Fuchs et al., 2008; Schatschneider, Wagner, & Crawford, 2008; Simmons et al., 2008; Speece & Case, 2001; Speece, Case, & Molloy, 2003; Vellutino et al., 2008) by focusing on students within an RtI model, by forming RD groups in word reading/text fluency and comprehension separately, by designating RD status after an instructional break, and by including descriptive analyses of classification rates for ELs. Specifying deficits in WR-F and C-V separately and in early 3rd grade effectively identified LEPRs, whose deficits may occur only in specific constructs (Catts et al., 2012). Indeed, 20% of RD students from Cohort A (4 in WR-F; 4 in C-V; 1 in combined) would not have been identified in the beginning of 2nd grade using identical RD classification procedures; approximately 11% from Cohort B (1 in WR-F and 3 in C-V) would not have been identified. In sum, results of the present study reflect the ability of early reading measures to

predict RD status for a diverse group of beginning 3rd grade students who showed early reading deficits, those who showed late-emerging deficits, and those who never showed deficits.

Our first goal was to identify 1st and 2nd grade reading measures useful for predicting RD in WR-F and C-V in 3rd grade. When included together in a prediction model for RD in WR-F, 1st grade WIF, and 2nd grade ORF and WRMT PC correctly classified 88.9% of RD and 86.2% of average readers. That 1st grade WIF may be useful in the prediction of RD is supported by earlier work, where WIF showed promise as a screening tool (Compton et al., 2006; 2010; Fuchs et al., 2008; Speece et al., 2011; Zumeta et al., 2012) and gauge of intervention responsiveness (Fuchs, Mock, Morgan, & Young, 2003). Results of prior studies are extended here, where WIF improved the prediction of RD beyond the end of 2nd grade, and where classifications were within an RtI framework with a diverse sample. The classification power of the isolated WIF/33rd percentile combination was especially impressive, considering WIF was administered 16 months before RD designations were made. Under this combination, 82% of RD (including 3 of 5 LEPRs) and 76.1% of average readers were correctly classified in Cohort A, with comparable rates in Cohort B.

In addition to identifying most students who would qualify as RD or average despite access to an additional year of intervention (i.e. persistent non-responders, O'Connor & Klingner, 2010), WIF identified students who, without intervention, might have been designated as RD in 3rd grade. These students received RD (mis)classifications by WIF (i.e. 1-specificity), but also received intervention in 2nd grade; thus, it is possible that some may have been identified as RD in 3rd grade without intervention. Although using single measures to refer students for more intensive intervention or special education does not reflect ideal practice (Denton, 2012), these results suggest that a WIF/33rd percentile combination, when estimated in spring of 1st

grade, may be a valuable part of more comprehensive models for identifying children at-risk for reading delays. Also, it is important to note that growth estimates for WIF were not explored here, but have been found useful for identifying children with persistent reading deficits at the end 1st (Speece et al., 2011) and 2nd (Compton et al., 2006) grade; future research might explore effects of WIF growth on RD classification made after 2nd grade.

Interestingly, fall-of-second grade WRMT PC raw scores were powerful predictors of 3rd grade RD group membership within the WR-F construct, even after accounting for effects of WIF and ORF. Relations between reading comprehension and fluency have been reported outside of RtI contexts with 4th grade students (Jenkins, Fuchs, van der Broek, Espin, & Deno, 2003) and 2nd grade students (Hudson et al., 2012), and might account for some of the predictive power of PC found here. Alternatively, that scores on the WRMT are somewhat reliant on a student's word reading skill (Keenan, Betjemann, & Olson, 2008) might also account for some of the predictive power. Nevertheless, PC showed promise as a model-based predictor.

PC also performed well as an isolated predictor. PC/25th correctly classified 79.3% of RD readers (including all LEPRs) and 81.6% of average readers in Cohort A, with comparable rates in Cohort B. Notably, more than 40% of misclassified average readers in Cohort A (30% in Cohort B) received intervention in 2nd grade, which might have altered their 3rd grade classifications. Thus like WIF, PC was able to identify persistent poor responders as well as students who might develop RD without immediate intervention.

Although the classification power of WRMT PC was noteworthy, measures of passage comprehension might be less attractive RtI indicators since they are often time consuming to administer compared to other measures (e.g. WIF, ORF) that provide equivalent or better classification rates. With these limitations in mind, evidence from the present study suggests that

within an RtI framework, scores on measures of passage comprehension with similar structures to the WRMT might play an important role in the prediction of RD in WR-F. Specifically, early 2nd grade PC scores may contribute to a comprehensive battery to assess responsiveness and predict RD in WR-F, or might be employed in "gated" screening procedures (Compton et al., 2010) for identifying at-risk students.

Finally and not surprisingly, ORF collected at the end of 2nd grade (and therefore in response to instruction or intervention in 2nd grade) was a model-based significant predictor of RD in WR-F[vii]; research consistently shows that disabled readers lag behind their average peers on measures of reading fluency. When paired with RtI criteria, however, the usefulness ORF varied. The ORF/final benchmark combination yielded excellent sensitivity and poor specificity for both cohorts, which is not too surprising since our whole-sample means for ORF fell below this cut-off. Nevertheless, that published benchmarks on ORF may be overly stringent indicators of poor response has been noted in other studies (Fuchs et al., 2004; O'Connor & Jenkins, 1999). When the final benchmark criterion was reduced to the highest score needed to capture all RD readers, specificity improved, resulting in acceptable classification rates for Cohort A and nearly acceptable rates for Cohort B. The new cut-off for risk also corresponded to the "at-risk" range on DIBELS benchmarks; therefore, sample-derived final benchmarks may be more useful for predicting RD in diverse populations, especially where ELs are prevalent.

In contrast to findings reported in Fuchs et al. (2004), ORF/low growth and ORF/DD criteria performed poorly when predicting RD in WR-F. This poor performance is especially troubling since the elapsed time between measure collection for prediction and measure collection for classification was short (i.e. the summer between 2nd and 3rd grade). In the present study, correct classification of RD students never exceeded 70%. Logically, many students

receiving access to intervention as needed might be expected to show reading growth. However, many children with adequate growth on ORF by end-of-second grade were designated RD by fall of 3rd grade, resulting in low sensitivity rates. The lapse in instruction over summer may have resulted in poorer performance, perhaps more reflective of response, on fall of 3rd grade assessments used to indicate RD. Another explanation might be gleaned from Burns and Senesac (2005), who used ORF scores to predict outcomes on a standardized reading assessment. The researchers found that although students demonstrated variable RtI, growth rates between response groups could not differentiate reading outcomes for 2nd graders with low end-of-year ORF scores. These findings suggest that growth commensurate with average readers may not be enough to shield poor readers from persistent reading problems, especially as reading demands change in 3rd and 4th grade. Therefore, researchers and practitioners must be careful when using commensurate growth or DD criteria to disqualify students for more intensive intervention, or to exit students from existing intervention. Even with similar growth, the gap between poor readers' skills and those of their typically developing peers might remain.

Differences in how reading outcomes were operationalized between this study and others may also account for the disagreement in the usefulness of DD as a response criterion. Outcomes in studies supporting DD were continuous scores on common reading assessments (Fuchs et al., 2004; Speece & Case, 2001). In these studies, DD students' outcome scores were significantly lower than scores for students who had good growth, but did not meet final achievement benchmarks. In the present study, students were classified as RD if their composite scores on WR-F fell 1.0 standard deviations below the sample mean. If DD classifies the *most* impaired readers on some skill, then the low sensitivity rates found in the present study may be explained in several ways. First, many students who were *not* DD on ORF still scored within the

RD range on WR-F, perhaps because the RD cut-off did not stringently identify only the most impaired readers. Also, some students who *were* DD on ORF had high scores on other measures of reading (e.g. spelling, word identification), which resulted in composites that did not meet the RD cut-off. Each of these scenarios may have led to poor classification accuracy.

Additionally, because DD may only identify students who improve the least during the classification years, the DD criterion may overlook late-emerging poor readers if used as the sole criterion to identify poor response to instruction. Readers with late-emerging RD tend to have lower scores than their peers in early grades, but their scores are not distinct enough to cause alarm (Catts et al., 2012; Compton et al., 2008). Therefore, although the ORF/DD combination might be useful for identifying the most impaired readers on ORF, many of whom will eventually develop RD, the combination might overlook students with less severe RD, over select students who struggle mostly in just one area of reading, and miss students who develop late-emerging RD. To alleviate some these drawbacks, DD might be applied to a combination of reading measures (see Fuchs et al., 2008), and the cut-point associated with growth and final achievement might be adjusted.

In addition to predicting RD in WR-F, we were interested in predicting RD in C-V. Logistic regression models identified 1st and 2nd grade TOLD and 2nd grade ORF as useful predictors of RD in C-V. When used in combination, these measures correctly classified 84.4% of RD readers and 76.7% of average readers in Cohort A. These rates fall short of field standards by approximately 5% each. Of the 41 over-classified readers, 22 (53.7%) never received intervention; thus they achieved "average" status without additional support. In many school contexts, providing intervention to 22 students who would have fared well without it unnecessarily depletes school resources that could have been used to improve instruction or

intervention for truly at-risk students. Additionally, that 5 students with RD were overlooked suggests need for future research into additional early indicators of reading comprehension-based disability. Early measures of reading comprehension, vocabulary, and fluency may be insufficient to accurately classify all students who will continue to struggle with higher-order reading skills. Measures of cognitive skill (e.g. oral language, nonverbal cognitive skill) may improve prediction (Catts et al., 2012).

Although the logistic regression model for C-V identified most students with RD, no stand-alone measure/criterion combination adequately predicted RD. The best sensitivity rates were for the ORF/final benchmark combination; however, classification of average readers was unacceptable. In contrast to results under the WR-F model, ORF fared poorly when paired with percentile-cut criteria. The highest sensitivity rate across cohorts was 78.6%, with specificity of 76.3%. First and 2nd grade TOLD/percentile combinations also produced low classification rates for each cohort; no more than 80% of RD readers were correctly classified.

The limited ability of ORF to adequately predict comprehension skill may be explained by the overall low language skill of our sample. Compared to national averages, students in our sample (NESs and ELs combined) demonstrated vocabulary skill deficits of almost 1 standard deviation below the national average on the PPVT and TOLD. In light of Crosson and Lesaux's (2010) work, text reading fluency skill of students in the present study (approximately half of whom were ELs) might not have facilitated reading comprehension due to their low language skills overall. In their study, Crosson and Lesaux (2010) reported that oral language skill (listening comprehension and to a lesser extent, vocabulary) moderated relations between text reading fluency and reading comprehension for Spanish-speaking 5th grade ELs. Text reading fluency was highly and positively associated with reading comprehension for children who also

demonstrated adequate oral language skills; for children with poorly developed listening comprehension, reading comprehension was low regardless of word reading skill and fluency. The results of the present study agree; for our students, who on average demonstrated poor language skill, text reading fluency was not an adequate indicator of future reading comprehension.

Since classification rates for single measure/criteria combinations were poor overall, we felt exploring prediction rates for LEPRs in C-V would be meaningless. Poor prediction power and our inability to meaningfully predict RD in C-V for LEPRs is certainly a limitation to this study. Nevertheless, we suggest that a model-based approach to identifying RD in C-V may be required to identify students who will continue to struggle with the complex skills involved in reading comprehension. Future research might explore the utility of measures not used here (e.g. measures of cognitive and oral language skills required for reading comprehension) to predict difficulty in reading comprehension via model-based and single-predictor approaches, for students who show early and later risk.

Classifying ELs within RtI

Logistic regressions for RD in each construct correctly classified NESs and ELs alike. Overall, members of each language group comprised misclassified average and RD reader groups within each construct and across cohorts, with no apparent systematically disproportionate misclassifications. An exception may be within the C-V model for Cohort B. Within that model, all misclassified readers with RD were ELs. Considering the work of Crosson and Lesaux (2010) described earlier, adding a measure of listening comprehension to our model may have resulted in better classification for ELs specifically. Overall, although

evidence is preliminary, our results suggest that model-based approaches for predicting RD in WR-F and C-V may apply similarly to NESs and ELs.

ELs were not disproportionately represented in misclassifications of RD and average readers on WRMT PC/25th percentile, which further supports this combination as an RtI indicator in the fall of 2nd grade within diverse populations. Conversely, promising 1st grade word reading and 2nd grade text fluency measure/criteria combinations may have consistently overlooked ELs with RD in WR-F. Though there did not appear to be disproportionate representation of ELs in the groups of average readers over-classified by WIF/33rd percentile in 1st grade, all misclassified RD readers across cohorts were ELs. This disproportionate misclassification is difficult to explain, especially since word reading fluency and text reading fluency are closely related for ELs over time (Crosson & Lesaux, 2010). Measurement error may have accounted for misclassifications under WIF, especially since scores for most misclassified ELs fell within 8 words of the WIF cut-off; however, scores for numerous correctly classified ELs and NESs were also borderline. The scores on 2nd grade ORF for most overlooked ELs fell well below the cut-off for risk, whereas ORF scores for correctly classified students with borderline WIF scores were comparatively higher. For the group of misclassified ELs, near average word reading fluency in 1st grade did not seem to translate to average text reading fluency in 2nd grade. Additional research might further explore the efficacy of word reading and text fluency measures for classifying ELs as RD within RtI contexts.

Similar to WIF, 7 of 9 (more than 75%) misclassified RD readers under the end-of 2nd grade ORF/25th combination were ELs, though we could not identify any systematic reasons for these misclassifications. Many misclassified students demonstrated drops in ORF over summer, as did many correctly classified students. It is possible that ELs had fewer opportunities over the

summer to read in English, resulting in poorer scores after break. Alternatively, although the intervention provided to ELs in 2nd grade focused on reading fluency and comprehension, it may not have adequately addressed their fluency needs. Finally, the disproportionate misclassification may have been due to error, but additional research is needed to rule-out substantively meaningful explanations.

Models vs. Single Measures

Support of model-based approaches (as opposed to single-measure approaches) for identifying children's response to instruction is strong in the literature (Compton et al., 2010; O'Connor & Jenkins, 1999). The present study supports this notion, especially for predicting RD in C-V and when RtI for ELs is explored. Indeed, classification rates were consistently higher within logistic regression models compared to most single-measure predictors, and the models did not appear to classify ELs nor NESs at disproportionate rates. It is interesting to note, however, that some single measure/criteria combinations classified RD and average readers within the WR-F construct with classification rates comparable to model-based approaches. Disregarding adjusted classification rates, WIF/33rd yielded sensitivity and specificity rates that fell short of the model-based rates by 4% and 13%, respectively, within Cohort A. Within Cohort B, sensitivity rates for the WIF/33rd percentile combination exceeded those of the model-based approach by 8%, though specificity fell below that of the model approach by approximately 13%. PC/25th percentile and ORF/25th also produced rates similar to the model-based rates. Overall, any of these measures may be viable substitutes for a model-based approach to RtI classification and/or for prediction of WR-F skill in 3rd grade. Though using single measure approaches as the only determinants of responsiveness or RD status may not be

ideal, these approaches are promising and may be a good first step for exploring student response.

Replication

A persistent need in intervention research involves model replication. The replication of regression models and measure/criteria combinations used for RD classification in the present study was impressive. It is important to note that students in each of the cohorts received instruction in similar environments (same schools, teachers, access to Tier II, and Tier II content when applicable), but differed in that students in Cohort B entered 1st grade the year following those in Cohort A. Therefore, replication was conducted on a sample very similar to the descriptive sample. Indeed, differences between cohorts in sensitivity and specificity indices for logistic regressions were small, with differences ranging from less than 1% to 7%. Differences in classification indices across Cohorts A and B were also small for measure/criteria combinations deemed useful in the prediction of RD, and fell between 1% and 5%. That the models developed with the original sample of students performed equally well when applied to a second sample provides support for many conclusions drawn in this study.

Limitations

Some features of this study that act as extensions to the literature may also serve as limitations. First, participants in this study received *access* to a high-quality Tier II intervention from 1st to 3rd grade. Some participants qualified for and received such intervention while others did not qualify for, and therefore did not receive, intervention. Yet, all students' scores were used in model-based and single-measure based predictions. This feature of the current study extends previous research in that early predictors of distal RD were collected under the assumption that students were receiving access to the most appropriate instruction available – the

case where a 1st grade poor reader has low scores due to inappropriate access to instruction is less likely. Although poor performance from participants in this study may still be a function of the instructional environment (i.e. even Tier II is not intensive enough to meet their needs), it is equally likely that poor responders in this study were truly at risk for later RD. That being said, results of this study hinge on the availability of intervention for struggling students in 1st and 2nd grade. Model-based and singular predictors of RD that showed promise in this study may be less powerful or inappropriate to use in studies or settings where children do not have fluid access to secondary intervention.

We also made a decision concerning cut-points that may affect interpretation and generalizability of results. We used a cut-off of 1.0 standard deviations below a locally normed mean to identify RD readers in this study, which seemed reasonable since sample-referenced cut-points are commonly used in studies where groups of RD and average readers are dichotomized based on some outcome or group of outcomes (e.g. Catts et al., 2012; Speece, Schatschneider, Silverman, Case, Cooper, & Jacobs, 2011). Also, we thought the sample prevalence rates for each of the WR-F (sample prevalence = 13% Cohort A; 14% Cohort B) and C-V (sample prevalence = 14% Cohort A; 15% Cohort B) constructs were reasonable; these rates mirrored those of other studies using a 1.0 standard deviation cut-off to define RD (e.g. sample prevalence of RD was 15% in Compton et al., 2012). Nevertheless, the cut-off used to identify RD and average readers in this study can be adjusted to meet the needs of specific contexts.

Additionally, although previous empirical research and analyses in the present study supported a two-factor structure to describe reading skill in 3rd grade, the reliability of each of the WR-F and C-V constructs should be tested on independent samples. Specifically, alternative measures of reading comprehension and vocabulary might be explored to create a construct that

describes skill in reading comprehension for elementary-aged students. Although average scores of 3rd grade RD readers on PPVT and WC raw were well below national averages and scores on the WRMT PC were about .75 SD below national norms, the utility of the C-V construct in the present study was low. Identifying children with deficits in specific areas of reading is important, especially when disability classifications are made later in schooling and when ELs, who may have strong decoding and fluency skill paired with poor comprehension and vocabulary knowledge are considered for intervention. Therefore, future research might explore measures of vocabulary and comprehension that better indicate RD in C-V alone.

Lastly, participants in this study attended high poverty schools and many were ELs; however, instruction provided in Tier I and Tier II may have been of higher quality than that offered in other similar settings. General education teachers in this study received 120 hours of professional development in the delivery of best-practice reading instruction. Also, many Tier II tutors were graduate students or former special education teachers, and most had experience working with children at-risk for reading delays. Thus, the Tier I and II instructional environments in this study may not reflect those in other low-income, high-minority schools. The extent to which results gained from this study apply to typical low-income instructional settings should be explored.

Practical Implications and Future Directions

Response-to-intervention is supported as an early intervention framework, but sparse research supports its role as an LD identification tool. Results from this study and others show that groups of good and poor responders vary depending on measures and criteria used to classify children into groups. Furthermore, prediction accuracy of distal RD based on early intervention measures hinges on the choice of predictors, criteria, and on the definition of RD.

Even with these considerations, some early reading measure/criteria combinations show promise as indicators of future RD for diverse students, at least in the areas of word reading and reading fluency.

Additional efforts should focus on identification of students who will show persistent difficulty in reading comprehension. Early text reading measures may do little to differentiate between students who will and will not become proficient comprehenders of text, especially for culturally and linguistically diverse students. According to results from the present study, multivariate methods may be needed to adequately identify children who will struggle with comprehension; thus, additional teacher training and time may be required to effectively translate such approaches into classroom practice. It also may be especially important to consider students' oral language skill when screening linguistically diverse children for intervention; measures of text reading fluency or vocabulary used alone may obscure teachers' understanding of these students' comprehension needs. Lastly, further research is required to determine whether children with late-emerging RD can be consistently identified using early screening measures. Failing to identify late-emerging poor readers before severe deficits develop will perpetuate the "wait-to-fail" cycle for these students, specifically.

References

Barth, A.E., Stuebing, K.K., Anthony, J.L., Denton, C.A., Mathes, P.G., Fletcher, J.M. & Francis, D.J. (2008). Agreement among response to intervention criteria for identifying responder status. *Learning and Individual Differences,18,* 296-307.

Beck, I.L., McKeown, M.G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. New York, NY: The Guilford Press.

Burns, M.K., & Sensac, B.V. (2005). Comparison of dual discrepancy criteria to assess response to intervention. *Journal of School Psychology, 43,* 393-406.

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities, 44*(5), 431-443.

Catts, H.W., Compton, D., Tomblin, J.B., & Sittner Bridges, M. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology, 104*(1), 166-181.

Chall, J.S., Jacobs, V.A., & Baldwin, L.E. (1990). *The reading crisis: Why poor children fall behind.* Cambridge, MA: Harvard University Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/ correlational analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Cooper, D.J., Pikulski, J.J., Ackerman, P.A., Au, K.H., Chard, D.J., Garcia, G.G., Goldenberg, C.N… Vogt, M. (2003). *Houghton Mifflin Reading.* Boston, MA: Houghton Mifflin Company.

Compton, D.L., Fuchs, D., Fuchs, L.S., Bouthon, B., Gilbert, J., Barquero, L.A., Crouch, R.C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327-340.

Compton, D.L., Fuchs, D., Fuchs, L.S, Bryant, J. (2006). Selecting at-risk reading in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394-409.

Compton, D.L., Fuchs, D., Fuchs, L.S., Elleman, A., & Gilbert, J.K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences, 18*(3), 329-337.

Compton, D.L., Gilbert, J.K., Jenkins, J.R., Fuchs, D., Fuchs, L.S, Cho, E., Barquero, L.A., Bouton, B.D. (2012). Accelerating chronically unresponsive children to tier 3 instruction:

What level of data is necessary to ensure selection accuracy? *Journal of Learning Disabilities, 45*(3), 204-216.

Crosson, A.C. & Lesaux, N.K. (2009). Revisiting assumptions about the relationship of fluent reading to comprehension: Spanish-speakers' text-reading fluency in English. *Reading and Writing: An Interdisciplinary Journal, 23,* 475-494.

Denton, C.A. (2012). Response to intervention for reading disabilities in the primary grades: Some answers and lingering questions. *Journal of Learning Disabilities, 45*(3), 232-243.

Dunn, L. M., & Dunn, D. M. (1997). *The Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Services, Inc.

Ehri, L.C. (1998). Grapheme–phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy.* (pp. 3–40). Mahwah, NJ: Erlbaum.

Francis, D.J., Fletcher, J.M., Stuebing, K.K., Reid Lyon, G., Shaywitz, B.A., & Shaywitz, S.E. (2005). Psychometric approaches to the identification of LD: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38*(2), 98-108.

Fuchs, D. & Fuchs, L.S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1).

Fuchs, D., Compton, D.L, Fuchs, L.S., Bryant, J., & Davis, N.G. (2008). Making "secondary intervention" work in a three-tier responsiveness-to-intervention model: findings from the first-grade longitudinal reading study of the National Research Center on Learning Disabilities. *Reading and Writing, 21,* 413-436.

Fuchs, D., Fuchs, L.S., & Compton, D.L. (2004). Identifying reading disabilities by responsiveness-to-intervention: Specifying measures and criteria. *Learning Disability Quarterly, 27*(4), 216-227.

Fuchs, D., Mock, D., Morgan, P.L., & Young, C.L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*(3), 157-171.

Good, R.H., & Kaminski, R.A. (2003). *Dynamic Indicators of Basic Early Literacy Skills 6th Edition* (6th ed.). Longmont, CO: Sopris West Educational Services.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257-288.

Hammill, D.D. & Newcomer, P.L. (2008). *Test of Language Development* (4th ed.). Austin, TX: Pro-Ed, Inc.

Hart, B. & Risley, T.R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H. Brookes Publishing.

Hudson, R.F. (2012). Fluency Problems: When, Why, and How to Intervene. In R.E. O'Connor & P. Vadasy (Eds.). *Handbook of Reading Interventions (pp. 169-199).* New York, NY: The Guilford Press.

Hudson, R.F., Torgesen, J.K., Lane, H.B., & Turner, S.J. (2012). Relations among reading skills and sub-skills and text-level reading proficiency in developing readers. *Reading and Writing, 25*(2), 483-507.

Jenkins, J. R., & O'Connor, R. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of Learning Disabilities*. (pps. 99-149). Hillsdale, NJ: Erlbaum.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, *95*(4), 719-729.

Klingner, J.K., Vaughn, S., & Boardman, A. (2007). Teaching Reading Comprehension to Students with Learning Difficulties. In K. Harris & S. Graham (Eds.). *What Works for Special-Needs Learners.* New York, NY: The Guilford Press.

Larsen. S., Hammill, D.D., & Moats, L. (1999). *Test of Written Spelling: Examiner's Manual*. Austin, TX: Pro-Ed, Inc.

Linan-Thompson, S. & Vaughn, S. (2003) *Research-Based Methods of Reading Instruction for English Language Learners Grades K-4.* Alexandria, VA: ASDC.

Mancilla-Martinez, J. & Lesaux, N. K. (2011). The gap between Spanish speakers' word reading and word knowledge: A longitudinal study. *Child Development, 82*(5), 1544–1560.

Mancilla-Martinez, J. & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology, 102,* 701–711.

Mellard, D.F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research & Practice, 24*(4), 186-195.

Morris, D., Trathen,W., Lomax, R.G., Perney, J., Kucan, L., Frye, E. M., …Schlagal, R. (2012). Modeling aspects of print-processing skill: implications for reading assessment. *Reading and Writing, 25*(1), 189-215.

Newcomer, P.L. & Hammill, D.D. (1997). *Test of Language Development* (3rd ed.). Austin, TX: Pro-Ed, Inc.

O'Connor, R. E. (2000). Increasing the intensity of intervention in kindergarten and first grade. *Learning Disabilities Research and Practice, 15*(1), 43-54.

O'Connor, R.E. (2007). Teaching Word Recognition: Effective Strategies for Students with Learning Difficulties. In K. Harris and S. Graham (Eds.). *What Works for Special-Needs Learners.* New York, NY: The Guilford Press.

O'Connor, R. E., Fulmer, D., Harty, K. R., & Bell, K. M. (2005). Layers of reading intervention in kindergarten through third grade: Changes in teaching and student outcomes. *Journal of Learning Disabilities, 38*(5), 440-455. doi: 10.1177/00222194050380050701

O'Connor, R. E. & Jenkins, J.R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3*(2), 159-197. doi: 10.1207/s1532799xssr0302_4

O'Connor, R.E. & Klingner, J. (2010). Poor responders in RTI. *Theory into Practice, 49*(4), 297-304.

Raudenbush, S.W. & Byrk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S.W., Byrk, A.S., Cheong, Y.F, Congdon, R. T., & Toit, M. (2011). *HLM 7: Linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Raykov, T. & Marcoulides, G. (2008). *An introduction to applied multivariate analysis.* New York, NY: Taylor and Francis Group.

Schnatschneider, C., Wagner, R., & Crawford, E. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18*, 308-315.

Schwanenflugel, P. J., Meisinger, E.B., Wisenbaker, J.M., Kuhn, M.R., Strauss, G.P., & Morris, R.D. (2006). Becoming a fluent and automatic reading in the early elementary school years. *Reading Research Quarterly, 41*(4), 496-522.

Simmons, D.C., Coyne, M.D., Kwok, O., McDonagh, S., Harn, B.A., & Kame'enui, E.J. (2008). Indexing response to intervention: A longitudinal study of reading risk from kindergarten through third grade. *Journal of Learning Disabilities, 41*(2), 158-173.

Simmons, D.C., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Wilson, V. , … Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness, 3*(2), 121-156.

Speece, D.L. & Case, L.P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*(4), 735-749.

Speece, D.L., Case, L.P., & Molloy, D.E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research &Practice, 18*(3), 147-156.

Speece, D.L., Schatscneider, C., Silverman, R., Case, L, Cooper, D.H., & Jacobs, D.M. (2011). Identification of reading problems in first grade within a response-to-intervention framework. *The Elementary School Journal, 111*(4), 585-607.

Spicer, J. (2005). *Making sense of multivariate data analysis.* Thousand Oaks, CA: Sage Publications.

Torgesen, J.K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice, 15*(1), 55-64.

Vadasy, P.F., Jenkins, J.R., Antil, L.R., Wayne, S.K., & O'Connor, R.E. (1997). The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers. *Learning Disability Quarterly, 20,* 126-139.

Vadasy, P. F., Sanders, E.A., & Tudor, S. (2007). Effectiveness of paraeducator-supplemented individual instruction: Beyond basic decoding skills. Journal of *Learning Disabilities, 40,* 508-525.

Vadasy, P.F., Wayne, S. K., O'Connor, R.E., Jenkins, J.R., Pool, K., Firebaugh, M., & Peyton, J. (2005). *Sound Partners: A tutoring program in phonics-based early reading.* Longmont, CO: Sopris West.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69*(4), 391-409.

Vellutino, F.R., Scanlon, D.M., & Lyon, R.G. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ-achievement discrepancy definition of reading disability. *Journal of Learning Disabilities, 33*(3), 223-238.

Vellutino, F.R., Scanlon, D.M., Sipay, E.R., Small, S.G., Chen, R., Pratt, A., & Denckla, M.B. (1996). Cognitive profiles of difficult to remediate and readily remediated poor readers: Early interventions as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*(4), 601-638.

Vellutino, F.R., Scanlon, D.M., Zhang, H., & Schatschneider, C. (2008). Using response to kindergarten and first grade intervention to identify children at-risk for long-term reading difficulties. *Reading and Writing, 21,* 437-480.

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. Scientific *Studies of Reading, 15*(1), 8-25.

Waesche, J.S., Schatschneider, C., Maner, J.K., Ahmed, Y., & Wagner, R. (2011). Examining agreement and longitudinal stability among traditional and RTI-based definitions of reading disability using the affected-status agreement statistic. *Journal of Learning Disabilities, 44*(3), 296-307.

Wanzek, J. & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*(4), 541-561.

Weiser, B. & Mathes, P. (2011). Using encoding instruction to improve the reading and spelling performances of elementary students at risk for literacy difficulties: A best-evidence synthesis. *Review of Educational Research, 81*(2), 170-200.

Wiederholt, J.L., & Bryant, B.R. (2001). *Gray Oral Reading Tests* (4th ed.). Austin, TX: Pro-Ed.

Woodcock, R. (1998). *Woodcock Reading Mastery Test–Revised/Normative Update*. Circle Pines, MN: American Guidance Service.

Zumeta, R.O., Compton, D.L., & Fuchs, L.S. (2012). Using word identification fluency to monitor first-grade reading development. *Exceptional Children, 78*(2), 201-220.

Tables

Table 1
Descriptive Statistics by Cohort

	Cohort A (0708 1st) *n* (proportion)	Cohort B (2008-2009 1st) *n* (proportion)	
Total N	219	168	
Gender (Male)	108 (.49)	94 (.56)	
Ethnicity			
African American	26 (.12)	12 (.07)	
Hispanic	156 (.71)	130 (.77)	
White	24 (.11)	17 (.10)	
Other	10 (.05)	6 (.04)	
Missing	3 (.01)	3 (.02)	
EL	110 (.50)	93 (.55)	
CELDT score Mean (SD)	3.05 (1.03)	3.0 (.90)	
	Mean (SD)	Mean (SD)	*p*-value
CST	319.87 (48.5)	334.93 (58.7)	
Outcome Measures			
3rd grade			
ORF	76.79 (26.1)	81.89 (29.6)	
TWS-4 raw*	10.16 (4.6)	10.68 (4.9)	
PPVT ss	86.2 (11.4)	87.95 (10.8)	
WRMT			
WID ss	102.6 (8.4)	104.8 (8.9)	
WA ss	103.58 (12.6)	106.93 (13.8)	
PC ss	100.93 (8.1)	102.45 (8.2)	
WC raw**	17.01 (5.8)	17.61 (6.4)	
Predictor Measures			
1st grade			
TOLDrv ss	7.3 (3.6)	8.3 (3.0)	.008
WIF	48.85 (22.5)	52.75 (24.5)	
ORF	52.37 (26.3)	55.04 (28.0)	
2nd grade			
TOLDrv ss	7.76 (2.7)	8.2 (2.8)	
ORF	89.38 (28.19)	80.52 (31.9)	
WRMT			
WID ss	107.31 (11.8)	109.46 (11.0)	
WA ss	105.63 (11.9)	110.31 (11.3)	<.001
PC ss	100.95 (9.6)	104.99 (9.0)	<.001
WC raw	10.75 (6.3)	9.64 (7.0)	

*Note. * The published raw score average on TWS-4 for 1st grade = 5. 2nd grade = 9 and 3rd grade = 13.* ** *The published raw score average on WRMT Word Comprehension (WC)for 3rd graders is 38,* p-values are reported for significant differences in means between cohorts; EL= English Language Learner; CST= California Standards Test; CELDT= California English Language Development Test (range 1-5); TWS-4= Test of Written Spelling- 4th Edition; PPVTss= Peabody Picture Vocabulary Test standard scored (normed

M=100; SD=15); TOLDrvss= Test of Language Development Relational Vocabulary Standard Score, 3rd Ed. (normed M=10; SD=3); ss standard score

Table 2
Tier II Intervention Inclusion and Exit Criteria

	1st Grade		2nd Grade	
	Inclusion Criteria	Exit Criteria	Inclusion Criteria	Exit Criteria
WIF	<8			
LNF	<45	> 35		
PSF	<30			
NWF	<25	> 30 fall > 50 winter, spring		
ORF (wcpm)		> 20 winter > 40 spring	<26 fall <52 winter <70 spring	> 44 fall > 68 winter > 90 spring

Note: WIF=Word Identification Fluency; LNF=Letter Naming Fluency; PSF= Phoneme Segmentation Fluency; NWF=Nonsense Word Fluency; ORF= Oral Reading Fluency; wcpm= words correct per minute. *Fall, winter, spring* refer to time points associated with inclusion and exit scores.

Table 3
Confirmatory factor analysis for 3rd grade outcomes by cohort

	Factor 1		Factor 2	
	Cohort A	Cohort B	Cohort A	Cohort B
ORF	.796	.834		
WRMT WID	.929	.935		
WRMT WA	.783	.804		
SPELLING	.850	.769		
WRMT WC			.607	.872
WRMT PC			.895	.745
PPVT			.383	.542

Model Fit statistics by cohort

	X^2	CFI	TLI	RMSEA (CI)
Cohort A	26.69 (p =.013)	.984	.974	.069 (.03, .11)
Cohort B	24.05 (p =.03)	.985	.976	.07 (.02, .12)

Table 4

Univariate Means, p-values, and effect sizes by construct and classification (Cohort A)

	Word Reading/Fluency Composite				Comprehension/Vocabulary Composite			
	RD	Average			RD	Average		
N	29	190			31	188		
Prevalence	13.2%	86.7%			14.2%	85.8%		
	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*
ORF	45.96 (4.3)	80.84 (1.7)	**34.88 (p<.001)**	.222/.53	55.4 (4.4)	79.8 (1.8)	24.2 (p<.001)	.131/.39
SPELLING	4.33 (0.77)	10.83 (0.3)	**6.49 (p<.001)**	.237/.56	6.3 (0.8)	10.57 (0.33)	4.26 (p<.001)	.11/.35
PPVT W	80.96 (2.2)	86.93 (0.86)	5.97 (p=.012)	.031/.18	72.62 (1.9)	88.41 (0.8)	**15.79 (p<.001)**	.233/.55
WRMT								
WID SS	92.63 (1.4)	103.87 (0.54)	**11.24 (p<.001)**	.225/.54	96.55 (1.4)	103.34 (0.59)	6.79 (p<.001)	.087/.31
WA SS	90.44 (2.0)	104.63 (0.81)	**14.19 (p<.001)**	.173/.46	95.86 (2.11)	103.88 (0.87)	8.072 (p=.001)	.06/.25
PC SS	90.85 (1.34)	102.28 (0.53)	11.42 (p<.001)	.242/.56	92.69 (1.3)	102.1 (0.55)	**9.41 (p<.001)**	.17/.45
WC Raw*	11.63 (1.03)	17.67 (0.41)	6.04 (p<.001)	.131/.39	9.66 (0.91)	18.08 (0.77)	**8.42 (p<.001)**	.269/.61
CST 3rd grade	274.96 (8.72)	326.83 (3.43)	51.87 (p<.001)	.133/.39	267.9 (8.1)	328.63 (3.3)	60.73 (p<.001)	.194/.49

*Notes: * The published raw score average on WC for 3rd graders is 3;* Bolded values are for construct-relevant measures; CST=California Standards Test; SS=Standard Score; SD=Standard Deviation; PPVT= Peabody Picture Vocabulary Test; W=winter; WRMT=Woodcock Reading Mastery Test; WID= Word Identification; WA= Word Attack; PC=Passage Comprehension; WC=Word Comprehension; Published Normative WID and WA Mean (SD) = 100(15); Cohen's F statistic adjusts for the upward bias in Eta-Squared. Cohen's F=$\sqrt{\frac{n^2}{1-n^2}}$

Table 5

Univariate Means, p-values, and effect sizes by construct and classification (Cohort B)

	Word Reading/Fluency Composite				Comprehension/Vocabulary Composite			
	RD	Average			RD	Average		
N	26	142			28	140		
Prevalence	15.5%	84.5%			16.7%	83.3%		
	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*
ORF	43.15 (4.8)	89.36 (2.0)	**46.21 (p<.001)**	.322/.69	51.79 (5.0)	87.9 (2.2)	36.11 (p<.001)	.208/.51
SPELLING	4.11 (0.77)	11.97 (0.33)	**7.86 (p<.001)**	.350/.73	5.3 (0.8)	11.8 (0.36)	6.5 (p<.001)	.246/.57
PPVTW	79.85 (2.0)	89.6 (0.85)	9.75 (p<.001)	.112/.36	76.79 (1.8)	90.18 (0.8)	**13.39 (p<.001)**	.221/.53
WRMT								
WID SS	93.12 (1.4)	107.09 (0.62)	**13.97 (p<.001)**	.325/.69	94.68 (1.5)	106.83 (0.65)	12.15 (p<.001)	.258/.59
WASS	92.65 (2.4)	109.74 (1.1)	**17.09 (p<.001)**	.201/.5	94.1 (2.4)	109.51 (1.1)	15.14 (p<.001)	.174/.46
PC SS	92.12 (1.34)	104.48 (0.58)	12.36 (p<.001)	.305/.66	91.43 (1.2)	104.7 (0.5)	**13.27 (p<.001)**	.367/.76
WC Raw*	9.81 (1.1)	19.1 (0.46)	9.29 (p<.001)	.281/.63	8.7 (.94)	19.4 (.42)	**10.7 (p<.001)**	.397/.81
CST 3rd grade	268.42 (12.2)	345.12 (4.7)	76.7 (p<.001)	.19/.48	270.25 (11.9)	344.81 (4.6)	74.56 (p<.001)	.187/.48

*Notes: * The published raw score average on WC for 3rd graders is 38.* Bolded values are for construct-relevant measures; CST=California Standards Test; SS=Standard Score; SD=Standard Deviation; PPVT= Peabody Picture Vocabulary Test; W=winter; WRMT=Woodcock Reading Mastery Test; WID= Word Identification; WA= Word Attack; PC=Passage Comprehension; WC=Word Comprehension; Published Normative WID and WA Mean (SD) = 100(15); Cohen's F statistic adjusts for the upward bias in Eta-Squared. Cohen's F=$\sqrt{\frac{n^2}{1-n^2}}$

Table 6
Stratified 3rd Grade Reader Classifications and Minutes of Tier II Intervention by Cohort

	n	Prevalence within sample	Prevalence within classification group	Received intervention Gr. 1* n (%)	Intervention Min Gr. 1 Mean (Range)**	Received intervention Gr. 2* n (%)	Intervention Min Gr. 2 Mean (Range)**
Cohort A	219						
Word Reading/ Fluency	14	6.4%	31.1%	2 (14.3%)	1086 (1023,1149)	9 (64.3%)	901.39 (370, 1551)
Comprehension/ Vocabulary	16	7.3%	35.5%	3 (18.8%)	1146.66 (910, 1335)	6 (37.5%)	817 (421, 1175)
Both	15	6.8%	33.3%	7 (46.7%)	1047.43 (585, 1665)	13 (86.7%)	1362.46 (890, 1995)
Total RD Readers***	**45**	**20.5%**	100%	12 (26.7%)	1078.67 (585, 1665)	28 (62.2%)	1087.38 (370, 1995)
Total Average Readers	**174**	**79.5%**	100%	6 (3.4%)	1193.58 (1015, 1540)	34 (19.5%)	1068.14 (65, 2023)
Cohort B	168						
Word Reading/ Fluency	8	4.8%	22.9%	7 (87.5%)	1100 (727, 1383)	7 (87.5%)	1340 (853, 2056)
Comprehension/ Vocabulary	9	5.4%	25.7%	5 (55.5%)	1158.4 (392, 1594)	5 (55.5%)	1236 (471, 2010)
Both	18	10.7%	51.4%	17 (94.4%)	1270.38 (609, 1696)	17 (94.4%)	1393.29 (550, 1775)
Total RD Readers ***	**35**	**20.9%**	100%	29 (82.8%)	1210.05 (392, 1696)	29 (82.8%)	1353.48 (471, 2167)
Total Average Readers	**133**	**79.2%**	100%	46 (34.6%)	899.06 (96, 1681)	42 (31.6%)	944.81 (75, 2080)

Note: * Percentages are out of total members within classification group ; ** Average and range reflect data from only students who received intervention within the corresponding group (Students with 0 intervention minutes were not included in averages or ranges); *** 9 (20%) RD readers from Cohort A and 4 (11.5%) from Cohort B would not have qualified as RD readers using 2nd grade data and did not qualify for intervention during 1st or 2nd grade; these students might be considered late-emerging poor readers.

Table 7
Pearson correlations for 3rd grade outcomes

	Cohort B							
	ORF	WRMT WID	WRMT WA	SPELLING	PPVT	WRMT PC	WRMT WC	CST
Cohort A								
ORF	1	.799	.639	.623	.353	.667	.544	.649
WRMT WID	.747	1	.746	.702	.492	.735	.636	.662
WRMT WA	.575	.722	1	.672	.344	.660	.557	.536
SPELLING	.685	.754	.710	1	.359	.666	.468	.610
PPVT	.185	.246	.230	.174	1	.453	.468	.513
WRMT PC	.602	.717	.594	.585	.298	1	.645	.662
WRMT WC	.375	.453	.400	.448	. 304	.538	1	.579
CST	.560	.572	.449	.546	.459	.603	.540	1

Note: Cohort A in rows and below the diagonal, Cohort B in columns and above the diagonal; all coefficients are significant at p<.001; all scores are raw scores.

Table 8
Pearson Correlations for 1st and 2nd grade predictors and 3rd grade outcomes

	1st WIF	1st ORF	2nd ORF	2nd WID	2nd WA	2nd WC	2nd PC	1st TOLD	2nd TOLD
1st WIF	1	.920	.846	.631	.650	.622	.631	.328	.367
1st ORF	.905	1	.865	.636	.687	.696	.689	.365	.404
2nd ORF	.789	.831	1	.672	.689	.626	.675	.288	.385
2nd WID	.541	.607	.549	1	.804	.554	.832	.447	.405
2nd WA	.573	.645	.518	.827	1	.551	.747	.302	.370
2nd WC	.598	.655	.567	.510	.576	1	.674	.505	.475
2nd PC	.570	.623	.554	.845	.783	.623	1	.463	.506
1st TOLD	**.140**	.208	.125(ns)	.252	.249	.247	.327	1	.545
2nd TOLD	.260	.331	.256	.372	.431	.426	.529	.454	1
Cohort A									
3rd ORF	.805	.827	.882	.562	.540	.546	.565	**.159**	.263
3rd WID	.637	.689	.657	.847	.800	.568	.793	.251	.344
3rd WA	.574	.613	.508	.713	.783	.542	.453	.237	.339
3rd Spell	.691	.704	.679	.632	.689	.649	.657	.246	.361
3rd WC	.369	.375	.364	.350	.387	.535	.453	.357	.460
3rd PC	.508	.555	.533	.760	.687	.520	.780	.265	.428
3rd PPVT	**.151**	.245	.247	.237	.280	.336	.372	.397	.506

Cohort B									
3rd ORF	.797	.820	.9	.583	.609	.584	.599	.283	.388
3rd WID	.646	.674	.732	.679	.696	.608	.726	.411	.473
3rd WA	.585	.607	.628	.675	.746	.503	.617	.266	.339
3rd Spell	.652	.695	.694	.614	.705	.600	.615	.375	.376
3rd WC	.442	.485	.537	.452	.487	.597	.596	.332	.493
3rd PC	.587	.605	.651	.753	.664	.609	.756	.406	.451
3rd PPVT	.226	.302	.318	.348	.296	.391	.427	.410	.409

Note. Top section: Cohort A correlations on rows and below diagonal; Cohort B correlations on columns and above diagonal;
Bottom section: Correlations of predictors with outcomes by cohort;
all ps<.01 except bolded = p<.05 and ns=not significant.

Table 9
Logistic regression models for word reading/fluency outcomes (Cohort A)

	b(SE)	Odds Ratio	*Best Predictors*[a]	Fit Criteria					
				AIC		BIC		-2 loglikelihood	
				Step	*Best Predictors*	*Step*	*Best Predictors*	*Step*	*Best Predictors*
Null				11004.8		110048.74		5489.40	
Model 1									
Step 1				3671.12	2037.66	3701.54	2054.56	1826.37	1013.83
Gender	-1.86**	.156	-1.82***						
	(.580)		(.57)						
Hispanic	-.354	.702							
	(.505)								
WIF1s	-.071*	.932	-.106***						
	(.033)		(.02)						
ORF1s	-.033	.986							
	(.027)								
Step 2				8092.57	3592.12	8183.57	3622.54	4019.29	1787.06
Gender	-2.34**	.097	-2.30**						
	(.875)		(.855)						
WIF1s	-.001	.999							
	(.037)								
ORF2s	-.109***	.897	-.101***						
	(.028)		(.017)						
WID 2ss	.024	1.02							
	(.054)								
WA 2ss	-.086	.917							
	(.049)								
PC2ss	-.158*	.854	-.21***						
	(.079)		(.052)						

Final Model			3592.12	3622.54	1787.06
Gender	-2.302**	.100			
	(.855)				
ORF2s	-.101***	.904			
	(.017)				
PC2ss	-.209***	.812			
	(.052)				

Note. a = beta coefficients(se) of model with only significant predictors at each step; s=spring time point; ss=standard score;1=1st grade; 2=2nd grade; WIF=Word Identification Fluency; ORF=Oral Reading Fluency; WID= Word Identification; WA= Word Attack; PC=Passage Comprehension; Final Model includes only significant predictors from steps 1 and 2 of Model 1.
**p<.05; **p<.01; ***p<.001*

Table 10
Logistic Regression Models for Comprehension/Vocabulary Outcomes (Cohort A)

	b(SE)	Odds Ratio	*Best Predictors*[a]	Fit Criteria					
				AIC		BIC		-2 Loglikelihood	
				Step	*Best Predictors*	*Step*	*Best Predictors*	*Step*	*Best Predictors*
Null Model				11575.68		11633.14		5770.84	
Model 1									
Step 1				4853.69	1328.51	4907.77	1345.41	2410.85	659.23
Gender	-1.09* (.45)	.33	-.99* (.41)						
Hispanic	-.09 (.57)	.92							
ELL1	.21 (.54)	1.23							
TOLD1ss	-.17** (.06)	.84	-.18*** (.05)						
ORF1	-.05 (.03)	.96							
WIF1	-.02 (.03)	.98							
Step 2				6979.105	3151.10	7070.36	3181.52	3462.55	1566.55
Gender	-1.12* (.47)	.33	-.92* (.47)						
TOLD1ss	-.06 (.08)	.94							
TOLD2ss	-.29* (.12)	.75	-.41*** (.09)						
ORF2s	-.02* (.01)	.97	-.04*** (.01)						
WC2raw	-.14 (.07)	.87							

PC2ss	.007 (.032)	.99			
Final Model			3151.10	3181.52	1566.55
Gender	-.923* (.47)	.40			
TOLD2ss	-.41*** (.09)	.67			
ORF2s	-.04*** (.011)	.96			

Note. a= beta coefficients(se)of model with only significant predictors at each step; s=spring time point; ss=standard score; raw=raw score (only raw scores were available for WC); 1=1st grade; 2=2nd grade; WIF=Word Identification Fluency; ORF=Oral Reading Fluency; WC=Word Comprehension; PC=Passage Comprehension; TOLD= Test of Oral Language Development; Final Model includes only significant predictors from Steps 1 and 2 of Model 1.

**p<.05; **p<.01; ***p<.001*

Table 11
Classification for RD in word reading/fluency by cohort

		Sensitivity		Specificity		Proportion LEPRs correctly classified	
	Cut-score* CA, CB	Cohort A	Cohort B	Cohort A	Cohort B	CA n=5	CB n=1
1st grade WIF**							
25th %tile	35, 36	75%	80.8%	81%	81.9%		
33rd %tile	38, 38	82.1%	88.5%	76.1%	76.1%	60%	100%
2nd grade PCss**							
25th %tile	95, 99	79.3%	80%	81.6%	83.6%	100%	0%
33rd %tile	97, 102	89.7%	88%	73.5%	74.3%		
2nd grade ORF							
Published Final benchmark	90	100%	100%	55.9%	43.2%		
Sample Final benchmark	75, 70	100%	100%	78.8%	69.1%		
Low Growth		37.9%	69.6%	89.4%	95%		
DD		40.7%	69.6%	95.1%	95%		
25th %tile	61, 62	85.2%	84.6%	83.2%	85.6%	80%	100%
33rd %tile	67, 67	93%	96.2%	75.5%	79.1%		

ROC Analysis* **	Cohort A			Cohort B		
	Cut score	AUC	Specificity	Cut score	AUC	Specificity
1st grade WIF	44.5	.856	62.1%	39.5	.905	74%
2nd grade PCss	98.5	.876	70%	103.5	.878	67.2%
2nd grade ORF	74.5	.925	80%	64.5	.948	81.8%

Note: * Students scoring below the cut score were poor responders; Cut scores for ORF are measured in words correct per minute.
**The rates reported here are conservative estimates since many students received intervention after 1st grade classifications were made or had reading deficits that developed after 2nd grade. See discussion for more information.
***ROC analysis is included to enable cross-study comparisons for the most promising combinations; sensitivity was set at ≥.90.
CA: Cohort A; CB = Cohort B; ss=standard scores; DD = Dual Discrepancy; LEPR = Late Emerging Poor Readers

Table 12
Classification for RD in comprehension/vocabulary by cohort

		Sensitivity		Specificity	
	Cut Score* CA, CB	Cohort A	Cohort B	Cohort A	Cohort B
1st grade TOLDss					
25th %tile	6, 6.5	50%	46.4%	71.8%	79.4%
33rd %tile	7, 8	56.3%	71.4%	66.3%	61%
2nd grade TOLDss					
25th %tile	7, 7	66.7%	53.6%	78.7%	83.6%
33rd %tile	8, 7	72.7%	64.3%	66.9%	72.9%
2nd grade ORF					
Published Final benchmark	90	81.3%	96.4%	54.2%	43.9%
Low Growth		24.2%	53.6%	87.5%	93.6%
DD		25%	53.6%	93.2%	93.6%
25th %tile	74, 62	46.9%	71.4%	78.2%	83.5%
33rd %tile	78, 67	59.4%	78.6%	70.4%	76.3%

Note: * Students scoring below the cut score were poor responders; Cut scores for ORF are measured in words correct per minute. CA: Cohort A; CB = Cohort B; ss=standard scores; DD = Dual Discrepancy; Prediction results for LEPRs are not tabled since no measure/criteria combination yielded adequate classification rates.

[i] We had substantive reasons to estimate a two-factor model with correlated factors; however we agreed with a reviewer who suggested it would be worthwhile to estimate additional models to determine whether our choice of model was the best fit for the data. Using data from Cohort A, we compared model fit of our original model with: 1) a one factor model, and 2) a two factor model with WRMT PC cross-loaded on the WR-F and C-V factors. Results indicated poor model fit for the one factor model (X2(14) = 50.36, p=.000; CFI = .951; TLI = .935; RMSEA = .11, CI = .08 to 14). The two factor model with WRMT PC cross-loaded fit the data as well as our original model (X2(12) = 22.28, p = .03; CFI = .988; TLI = .979; RMSEA = .06, CI = .02, .1). Although the model with cross-loadings indicated good model fit, we retained our original, theoretically driven model to represent reading constructs. This is because although scores on WRMT PC are known to rely on word reading to some degree (and therefore are expected to show relations with word reading and comprehension factors), WRMT PC is used practically as a measure of reading comprehension. We felt that either model was defensible, but the model without cross loadings allowed for a more straightforward interpretation of the factors.

[ii] Sample means and standard deviations for many 3rd grade assessments approximated published norms; for these, the bottom 16th percentile based on sample norms and published norms would result in nearly identical RD and average reader groups. The spelling and vocabulary measures were the exception; using published norms to indicate risk on these measures would have resulted in severely inflated RD groups, and likely would not represent realistic criteria to use in school-implemented RtI settings where many students are English Learners and come from impoverished backgrounds.

[iii] $e^{-.101} = .904;\ 100(.904 - 1) = -9.6.$

[iv] Chi Square statistics are not provided in Mplus; therefore, difference tests were conducted using the -2 log likelihood for the intercept-only model compared to prediction models with degrees of freedom equal to the number of independent variables. The null hypothesis states that addition of the predictors does not add to knowledge of group membership. Rejection of the null indicates model improvement. This test is analogous to a test of R-square change in step-wise regression analysis and is an appropriate alternative index of model fit (Spicer, 2005).

[v] Adjusted 1-specificity = (1-specificity)*(proportion of over-classified readers with no intervention in 2nd grade); for example, WIF/33rd percentile adjusted specificity is (1-.761)* (.295) = .07. Specificity = 1- .07 = .93.

[vi] by design, the final benchmark score was lowered to reflect the lowest ORF score that would still yield sensitivity of 100%

[vii] Exploring the ability of an end-of-second grade measure to predict RD when designations were made in fall-of-3rd grade may seem to lack educational relevance. Nevertheless, text-reading fluency measures receive considerable attention in response-to-intervention frameworks, and so we wanted to include in our prediction models and as an isolated predictor a measure of text-reading fluency collected after receipt of instruction and intervention in 2nd grade. Given the relatively poor performance of particular ORF/criterion combinations found in this study, we felt retaining ORF performance results and our interpretation of the results was important.